

EXHIBIT 2

Part 1



DEPARTMENT OF HEALTH & HUMAN SERVICES

National Institutes of Health
National Human Genome Research Institute
FOIA/PA Office, RKL 1, 4th Floor
6705 Rockledge Drive
Bethesda, MD 20892

February 8, 2023

Sarah M. Cork
Quinn Emanuel Urquhart & Sullivan LLP
630 Idaho Ave
Apt 304
Santa Monica, CA 90403

Re: FOIA Case Number: 59031

Dear Dr. Cork:

This is our final response to your Freedom of Information Act (FOIA) request addressed to the National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH), dated September 22, 2022 and received the same day. You requested all relevant records concerning project P50HG005550, including the full grant application and any drafts thereof, any grant- or pre-grant-related correspondence involving the Awardee, the Awardee's affiliates, or the Awardee Institute, and any related progress reports, invention reports, information disclosures, financial reports, and any other information concerning activities conducted in relation to this grant and application. This request also extends to any available data associated with this project, which is requested pursuant to OMB Revised Circular A110.

On September 30, you agreed to exclude data related to OMB Circular A110 Revised from the requested material via email. In an email on January 10, 2023, you agreed to accept our standard redactions and waive your rights to appeal these redactions.

Enclosed are 906 pages responsive to your request. This includes the funded application and updates, subsequent annual applications, a supplement application, progress reports, notices of award and their revisions, correspondence with NHGRI program staff, federal financial reports, financial carryover requests and approvals, invention reports, a publications report, and abstracts and presentations sent to program staff. It is Department of Health and Human Services (HHS) policy to expunge consultant information such as hourly rates, effort levels, eRA Commons user names, estimated costs, evaluative information, institutional base salary, pending support, personal information, priority scores, private sources of funding, reviewers' comments, signatures, and unpublished scientific information wherever they appear throughout the grant material. Summary statements are expunged of the priority score, direct costs recommended, evaluation, opinion and information pertaining to the budget recommendation. This information has been removed from the enclosed material.

Requesters who ask for grant applications usually want to receive only material that will help in understanding the process that led to the awards, or to improve their own methods of drafting grant applications. Requesters usually do not want material that applicants believe would harm them if released. We have found that the spirit of the FOIA can be enhanced through a spirit of cooperation among requesters of materials and those who submitted the materials.

Re: FOIA Case Number: 59031

In this instance, we asked the grantee for advice concerning patent rights and other confidential commercial or financial information and the material that we are furnishing reflects that advice. If you feel that materials have been omitted that should have been made available to you, please write to me and I will consult with the NIH Freedom of Information Officer. Please feel free to call me on 301-496-9737 for additional information or to inquire about your request.

If you are not satisfied with the processing and handling of this request, you may contact the NHGRI FOIA Public Liaison:

NHGRI FOIA Public Liaison

Marianne Manheim
Rockledge I, 4th Floor
Bethesda, MD 20892
301-496-9737 (phone)
301-402-3604 (fax)
marianne.manheim@nih.gov (email)

In certain circumstances, provisions of the FOIA and HHS FOIA Regulations allow us to recover part of the cost of responding to your request. Enclosed is an invoice for \$ 736.00 to cover the costs of responding to your request. Please note that NIH now accepts electronic payments. Instructions are included with the invoice.

Sincerely,
/S/
Valery Gheen
Government Information Specialist, NHGRI

Enclosures: 906 pages
Invoice

National Institutes of Health
Invoice of Fee for Freedom of Information Act(FOIA) Services



Building 31 Room 5B35
 9000 Rockville Pike
 Bethesda, MD 20892

Requester Details

Sarah Cork
 attorney
 Quinn Emanuel Urquhart & Sullivan LLP
 630 Idaho Ave
 Apt 304
 Santa Monica, CA 90403

Requester Invoice

Request No : 59031
 Invoice No : 00000002257
 Invoice Date : 02/08/2023
 Requester Name : Sarah Cork

Fee Item	Quantity	Unit (\$)	Extended (\$)	Charged (\$)
Review Costs				
Review Rate 2	16.00	\$46.00	\$736.00	\$736.00
Total Amount				\$736.00
Amount Paid				\$0.00
Balance Due				\$736.00

National Institutes of Health
Invoice of Fee for Freedom of Information Act(FOIA) Services

Request Description

Please see the attached letter.

Briefly, this request relates to records for Application No. 7849826 and Grant No. 1P50HG005550-01, titled "Causal Transcriptional Consequences of Human Genetic Variation" to George Church, which are responsive to Program Announcement PAR-08-094.

Thank you for your consideration, and please let me know if you need more information. (Date Range for Record Search: From 02/22/2008 To 07/31/2015)

Sub Requests

Default

Invoice Memo

Instructions: We can only accept online electronic payments.

Electronic Payments:

Automated Clearing House (ACH) electronic debit via Pay.gov:

For additional information, see <https://pay.gov>

If you need assistance using this site, please contact Pay.gov Customer Service at 1-800-624-1373. You may pay NIH FOIA Payments here: <https://www.pay.gov/public/form/start/38036554>.

Payment is due within 30 days from the date of this invoice. Interest will be charged after the due date.

Footer Note

Form Approved Through 11/30/2010

OMB No. 0925-0001

Department of Health and Human Services Public Health Services		PI: CHURCH, GEORGE M		Council: 01/2010	
12155092		Application MAY		1 P50 HG005550-01	
<small>Do not exceed character length restrictions indicated.</small>		Dual: MH		Received: 05/21/2009	
IRG: ZHG1 SRC(99)					
1. TITLE OF PROJECT (Do not exceed 81 characters, including spaces and punctuation.) "Causal Transcriptional Consequences of Human Genetic Variation"					
2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION <input type="checkbox"/> NO <input checked="" type="checkbox"/> YES					
Number: PAR-08-094		Title: Centers of Excellence in Genomics Science			
3. PROGRAM DIRECTOR/PRINCIPAL INVESTIGATOR			New Investigator <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		
3a. NAME (Last, first, middle) Church, George M.		3b. DEGREE(S) PhD.		3h. eRA Commons User Name eRA Commons User Name	
3c. POSITION TITLE Professor		3d. MAILING ADDRESS (Street, city, state, zip code) Harvard Medical School NRB 238 77 Avenue Louis Pasteur Boston, MA 02115			
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Genetics					
3f. MAJOR SUBDIVISION School of Medicine					
3g. TELEPHONE AND FAX (Area code, number and extension) TEL: 617-432-7562 FAX: 617-432-6513		E-MAIL ADDRESS: gmc@harvard.edu			
4. HUMAN SUBJECTS RESEARCH <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes		4a. Research Exempt If "Yes," Exemption No. <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes			
4b. Federal-Wide Assurance No. FWA00007071		4c. Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		4d. NIH-defined Phase III Clinical Trial <input type="checkbox"/> No <input type="checkbox"/> Yes	
5. VERTEBRATE ANIMALS <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes			5a. Animal Welfare Assurance No.		
6. DATES OF PROPOSED PERIOD OF SUPPORT (month, day, year—MM/DD/YY) From 4/1/10 Through 3/31/15		7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD 7a. Direct Costs (\$) 2,800,000		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT 7b. Total Costs (\$) 4,416,977 8a. Direct Costs (\$) 12,085,320 8b. Total Costs (\$) 20,062,090	
9. APPLICANT ORGANIZATION Name President and Fellows of Harvard College Address Harvard Medical School Sponsored Programs Administration 25 Shattuck Street Room 509 Boston, MA 02115		10. TYPE OF ORGANIZATION Public: <input type="checkbox"/> Federal <input type="checkbox"/> State <input type="checkbox"/> Local Private: <input checked="" type="checkbox"/> Private Nonprofit For-profit: <input type="checkbox"/> General <input type="checkbox"/> Small Business <input type="checkbox"/> Woman-owned <input type="checkbox"/> Socially and Economically Disadvantaged			
		11. ENTITY IDENTIFICATION NUMBER 1042103580C5 DUNS NO. 047006379 Cong. District 8th			
12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE Name Deborah Good Title Assoc Dir - Sponsored Programs Admin Address Harvard Medical School 25 Shattuck Street Boston, MA 02115 Tel: 617-432-1596 FAX: 617-432-2651 E-Mail: spa_award@hms.harvard.edu		13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION Name Deborah Good Title Assoc Dir - Sponsored Programs Admin Address Harvard Medical School 25 Shattuck Street Boston, MA 02115 Tel: 617-432-1596 FAX: 617-432-2651 E-Mail: spa_award@hms.harvard.edu			
14. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.		SIGNATURE OF OFFICIAL NAMED IN 13. (Do not ink "Do" signature not acceptable) Signature		DATE 5/19/09	

Program Director/Principal Investigator (Last, First, Middle):

Church, George M.

PROJECT SUMMARY (See instructions):

The Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV) will develop innovative and powerful genetic engineering methods and use them to identify genetic variations that causally control gene transcription levels. Genome Wide Association Studies (GWAS) find many variations associated with disease and other phenotypes, but the variations that may actually cause these conditions are hard to identify because nearby variations in the same haplotype blocks consistently co-occur with them in human populations, so that specifically causative ones cannot be distinguished. About 95% of GWAS variations are not in gene coding regions, and many of these presumably associate with altered gene expression levels. CTCHGV will identify the variations that directly control gene expression by engineering precise combinations of changes to gene regulatory regions that break down the haplotype blocks, allowing each variations' effect on gene expression to be discerned independently of the others. To perform this analysis, CTCHGV will extract ~100kbps gene regulatory regions from human cell samples, create precise variations in them in *E. coli*, and re-introduce the altered regions back into human cells, using zinc finger nucleases (ZFNs) to efficiently induce recombination. CTCHGV will target 1000 genes for this analysis (Aim 1), and will use human induced Pluripotent Stem cells (iPS) to study the effects of variations in diverse human cell types (Aim 2). To explore the effects of variations in complex human tissues, CTCHGV will develop methods of measuring gene expression at transcriptome-wide levels in many single cells, including *in situ* in structured tissues (Aim 3). Finally, CTCHGV will develop novel advanced technologies that integrate DNA sequencing and synthesis to construct thousands of large DNA constructs from oligonucleotides, that enable very precise targeting and highly efficient performance of ZFNs, and that enable cells to be sorted on the basis of morphology as well as fluorescence and labeling (Aim 4). CTCHGV will also develop direct oligo-mediated engineering of human cells, and create "marked allele" iPS that will enable easy ascertainment of complete exon distributions for many pairs of gene alleles in many cell types.

RELEVANCE (See instructions):

CTCHGV methods will yield precise knowledge of effects of human genetic variations on gene expression that will both refine and go beyond GWAS-derived associations between non-coding variations and disease. Powerful new CTCHGV genetic engineering methods will directly enable gene therapy. CTCHGV iPS and single-cell transcriptome technologies will increase understanding of diverse and complex human tissues.

PROJECT/PERFORMANCE SITE(S) (if additional space is needed, use Project/Performance Site Format Page)

Project/Performance Site Primary Location: PD/ PI: HMS : George M. Church			
Organizational Name: Harvard Medical School:			
DUNS: 047006379			
Street 1: 77 Avenue Louis Pasteur		Street 2: Rm 238	
City: Boston	County: Suffolk	State: MA	
Province:	Country: USA	Zip/Postal Code: 02115	
Project/Performance Site Congressional Districts: 8th			
Additional Project/Performance Site Location: Subcontract: UCSD: Kun Zhang			
Organizational Name: University of California			
DUNS: 80-435-5790			
Street 1: 9500 Gilman Drive		Street 2:	
City: La Jolla	County:	State: CA	
Province:	Country: USA	Zip/Postal Code: 92093	
Project/Performance Site Congressional Districts: 53			

Program Director/Principal Investigator (Last, First, Middle): Church, George M.
--

Additional Project/Performance Site Location: Subcontract: Children's Hospital : George Daley
--

Organizational Name: Children's Hospital Boston

DUNS: 076593722

Street 1: 300 Longwood Ave., KARP Res Bldg., 7 th fl.	Street 2:
--	-----------

City: Boston	County: Suffolk	State: MA
--------------	-----------------	-----------

Province:	Country: USA	Zip/Postal Code: 02115
-----------	--------------	------------------------

Project/Performance Site Congressional Districts: 8th

Additional Project/Performance Site Location: Subcontract: CCIB - MGH: Keith Joung

Organizational Name: Massachusetts General Hospital- CCIB

DUNS: 073130411

Street 1: 13 th Street	Street 2: Bldg 149 6 th fl
-----------------------------------	---------------------------------------

City: Charlestown	County: Suffolk	State: MA
-------------------	-----------------	-----------

Province:	Country:	Zip/Postal Code: 02114
-----------	----------	------------------------

Project/Performance Site Congressional Districts: MA-009
--

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

SENIOR/KEY PERSONNEL. See instructions. Use continuation pages as needed to provide the required information in the format shown below. Start with Program Director(s)/Principal Investigator(s). List all other senior/key personnel in alphabetical order, last name first.

Name	eRA Commons User Name	Organization	Role on Project
Church, George, PhD.	eRA Commons User Name	Harvard Medical School	PD/PI
Zhang, Kun, PhD.		UCSD	Co-I
Keith Joung, MD, PhD.		MGH	Co-I
Daley, George, MD, PhD.		Children's Hospital Boston	Co-I

OTHER SIGNIFICANT CONTRIBUTORS

Name	Organization	Role on Project
------	--------------	-----------------

Human Embryonic Stem Cells ☒ No ☐ Yes

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: <http://stemcells.nih.gov/research/registry/>. Use continuation pages as needed.

If a specific line cannot be referenced at this time, include a statement that one from the Registry will be used.

Cell Line

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

RESEARCH GRANT TABLE OF CONTENTS

	<i>Page Numbers</i>
Face Page	1
Description, Project/Performance Sites, Senior/Key Personnel, Other Significant Contributors, and Human Embryonic Stem Cells	2-4
Table of Contents	5
COMPOSITE BUDGET yr1: Detailed Budget for Initial Budget Period	6
COMPOSITE BUDGET: Budget for Entire Proposed Period of Support	7
MAIN CEGS BUDGET yr1: Detailed Budget for Initial Budget Period	8
MAIN CEGS BUDGET: Budget for Entire Proposed Period of Support	9
Budget Justification: MAIN	10-14
MAP BUDGET yr1: Detailed Budget for Initial Budget Period	15
MAP BUDGET: Budget for Entire Proposed Period of Support	16
Budget Justification: MAP	17
Consortium/Contractual Budgets: CHB, UCSD, MGH	18-32
 Biographical Sketch – Program Director/Principal Investigator (Not to exceed four pages each)	 33-36
Other Biographical Sketches (Not to exceed four pages each – See instructions)	37-46
Resources	47-51
Checklist	52-53
Research Plan	54-103
1. Introduction to Resubmission Application, if applicable (Not to exceed three pages.), or Introduction to	
2. Specific Aims	54
3. Background and Significance	56
4. Preliminary Results	60
5. Research Design	67-103
6. Bibliography and References	104
6. Management and Organization	117
7. Training Plan	123
8. Minority Action Plan	125
9. Data and Materials Dissemination Plan	130
10. Inclusion Enrollment Report (Renewal or Revision applications only)	132
11. Protection of Human Subjects	132
12. Inclusion of Women and Minorities	132
13. Targeted/Planned Enrollment Table	132
14. Inclusion of Children	132
15. Vertebrate Animals	132
16. Select Agent Research	132
17. Multiple PD/PI Leadership Plan	133
18. Resource Sharing Plan	133
19. Consortium/Contractual Arrangements	133
20. Letters of Support (e.g., Consultants)	135-141
Appendix (Five identical CDs.)	<div style="display: flex; align-items: center; justify-content: flex-end;"> <input style="width: 20px; height: 20px; margin-right: 10px;" type="checkbox"/> <div style="text-align: left;"> Check if Appendix is included </div> </div>

Composite budget yrr

Program Director/Principal Investigator (Last, First, Middle):

Church, George M.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY						FROM 4/1/10	THROUGH 3/31/11	
PERSONNEL (Applicant organization only)		Months Devoted to Project			DOLLAR AMOUNT REQUESTED (omit cents)			
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	INST.BASE SALARY	SALARY REQUESTED	FRINGE BENEFITS	TOTAL
Church main						703,149	194,131	897,281
Map						170,285	44,415	214,700
MGH: Joung						see	consortium	
Children's: Daley						see	consortium	
UCSD: Zhang						see	consortium	
SUBTOTALS →						873,434	238,546	1,111,980
CONSULTANT COSTS								
Main- consultants: \$83,000								83,000
EQUIPMENT (Itemize)								
Specialized equipment: \$500,000								500,000
SUPPLIES (Itemize by category)								
Main : \$325,000								
MAP : \$75,300								400,300
TRAVEL								
Main: \$18,000								
MAP: \$10,000								28,000
PATIENT CARE COSTS		INPATIENT						
		OUTPATIENT						
ALTERATIONS AND RENOVATIONS (Itemize by category)								
OTHER EXPENSES (Itemize by category)								
Main: \$344,041								344,041
CONSORTIUM/CONTRACTUAL COSTS					DIRECT COSTS		332,679	
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD (Item 7a, Face Page)								\$ 2,800,000
CONSORTIUM/CONTRACTUAL COSTS					F&A cost		221,013	
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD								\$ 3,021,013

Composite budget - 5 yrs

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

**BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY**

BUDGET CATEGORY TOTALS		INITIAL BUDGET PERIOD (from Form Page 4)	ADDITIONAL YEARS OF SUPPORT REQUESTED			
			2nd	3rd	4th	5th
PERSONNEL: <i>Salary and fringe benefits. Applicant organization only.</i>		1,111,981	1,153,570	1,154,028	1,148,958	1,161,608
CONSULTANT COSTS		83,000	83,000	83,000	83,000	83,000
EQUIPMENT		500,000				
SUPPLIES		400,300	398,816	408,408	416,601	408,840
TRAVEL		28,000	28,300	22,609	16,927	17,255
PATIENT CARE COSTS	INPATIENT					
	OUTPATIENT					
ALTERATIONS AND RENOVATIONS						
OTHER EXPENSES		344,041	309,380	304,625	306,839	301,268
CONSORTIUM/ CONTRACTUAL COSTS	DIRECT	332,679	335,215	344,139	353,268	362,670
SUBTOTAL DIRECT COSTS (Sum = Item 8a, Face Page)		2,800,000	2,308,280	2,316,808	2,325,592	2,334,640
CONSORTIUM/ CONTRACTUAL COSTS	F&A	221,013	234,535	241,199	247,515	254,020
TOTAL DIRECT COSTS		3,021,013	2,542,815	2,558,007	2,573,107	2,588,660
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD						\$ 13,283,602

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Equipment – 500K (specialized equipment), not part of our direct cost

Main - Harvard Y11

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY						FROM 4/1/10	THROUGH 3/31/11	
PERSONNEL (Applicant organization only)		Months Devoted to Project			INST. BASE SALARY	DOLLAR AMOUNT REQUESTED (omit cents)		
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths		SALARY REQUESTED	FRINGE BENEFITS	TOTAL
Church, George M.	PD/PI	EFFORT			Institutional Base Salary	70,812	18,836	89,648
John Aach	Software Eng					65,347	17,382	82,730
Richard Terry	Engineer					100,055	42,023	142,078
Yveta Masarova	Admin.					5,246	2,203	7,450
Sara Vassallo	Technician					47,418	25,890	73,309
Francois Vigneault	PostDoc					10,551	2,849	13,400
Dae Kim	PostDoc					29,543	7,977	37,519
Jin Billy Li	PostDoc					31,853	8,600	40,453
Jehyuk Lee	PostDoc					43,860	11,842	55,702
Morten Sommer	PostDoc					40,500	10,935	51,435
Michael Sismour	PostDoc					45,504	12,286	57,790
Tara Gianoulis	PostDoc					39,000	10,530	49,530
Daniel Levner	PostDoc					43,860	11,842	55,702
Sasha Wait Zaranek	PostDoc					40,500	10,935	51,435
Abraham Rosenbaum	Grad student					29,700		29,700
Madeleine Ball	Grad student					29,700		29,700
Xavier Rios	Grad student					29,700		29,700
SUBTOTALS						703,149	194,131	897,281
CONSULTANT COSTS: \$ 83,000								83,000
EQUIPMENT (Itemize): See justification specialized equipment : \$500,000								500,000
SUPPLIES (Itemize by category): Church lab: \$300,000 Computers: \$25,000								325,000
TRAVEL: \$18,000								18,000
PATIENT CARE COSTS		INPATIENT						
		OUTPATIENT						
ALTERATIONS AND RENOVATIONS (Itemize by category)								
OTHER EXPENSES (Itemize by category)								
Harvard Tuition: A. Rosenbaum, M. Price Ball: (2x5,070)=\$10,140 Xavier Rios: \$11,600								
Grad student fee: (4x3,000): \$12,000								
MIT Tuition: Uri Laserson (tuition & stipend: Institutional Base Salary): \$43,896								
Service contracts \$ 36,000								
Media Services & Glass washing: \$ 5,000								
Publications: \$10,000								
Sequencing: \$215,405								344,041
CONSORTIUM/CONTRACTUAL COSTS						DIRECT COSTS		
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD (Item 7a, Face Page)						\$ 2,167,322		
CONSORTIUM/CONTRACTUAL COSTS:						FACILITIES AND ADMIN COSTS:		
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD						\$ 2,167,322		

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

**BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY**

BUDGET CATEGORY TOTALS		INITIAL BUDGET PERIOD (from Form Page 4)	ADDITIONAL YEARS OF SUPPORT REQUESTED			
			2nd	3rd	4th	5th
PERSONNEL: <i>Salary and fringe benefits. Applicant organization only.</i>		897,281	933,149	927,714	916,575	922,973
CONSULTANT COSTS		83,000	83,000	83,000	83,000	83,000
EQUIPMENT		500,000				
SUPPLIES		325,000	321,257	328,523	334,318	324,089
TRAVEL		18,000	18,000	12,000	6,000	6,000
PATIENT CARE COSTS	INPATIENT					
	OUTPATIENT					
ALTERATIONS AND RENOVATIONS						
OTHER EXPENSES		344,041	309,380	304,625	306,839	301,268
CONSORTIUM/ CONTRACTUAL COSTS	DIRECT					
SUBTOTAL DIRECT COSTS (Sum = Item 8a, Face Page)		2,167,322	1,664,786	1,655,862	1,646,732	1,637,330
CONSORTIUM/ CONTRACTUAL COSTS	F&A					
TOTAL DIRECT COSTS		2,167,322	1,664,786	1,655,862	1,646,732	1,637,330
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD						\$ 8,772,032

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Equipment – 500K (specialized equipment), not part of our direct cost

/

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): Church, George M.

Budget Justification / yr 1 - HARVARD**Harvard Personnel:****General notes:**

Base salaries and fringe rates assume a Sept. 1st start date for the budget years and a July 1 start date for the fiscal year at Harvard Medical School. Fringe benefit rates for the current year are 26.6% for the Faculty, 27% for postdoctoral fellows, total fringe rates - 54.6% for non-exempt employees (research and computer technicians) 42% exempt employees (admin and professional).

"The staff salary expenses in this budget were calculated in accordance with the Harvard Treatment of Paid Absences portion of Section II of our rate agreement negotiated with the Department of Health and Human Services on *February 27, 2008*. Harvard will begin using a vacation accrual beginning July 1, 2007. Paid absences for vacation will no longer be claimed as direct charged on federal awards and regular salary will carry a vacation fringe to accrue earned vacation."

Postdoctoral salaries for Year 1 reflect either the rates recommended by NIH based on years of experience or, in some cases, the current salary level for certain individuals.

Church lab personnel:

George Church, PhD (Program Director/ Principal Investigator), effort on the project –
George Church is the PI in charge of the entire CEGS. He will oversee all project work, coordinate roles with collaborating investigators, and be the principal contact of the CEGS with NIH, external collaborators (including companies), scientific and IRB advisory boards, and other centers and institutions for which there is a need to coordinate activities (such as other CEGS).

John Aach, PhD (Software engineer), effort on the project –
John Aach is a Lecturer in the Church Lab in the HMS Department of Genetics. He will assist George Church in supervising the CEGS and will work on image and computational analysis and statistical methods for the Center.

Richard Terry, MS (Engineer, Church lab) effort on the project –
Richard is a senior engineer. Richard will conduct and oversee engineering research and device production. This also covers device development, installation, operations, maintenance, and budgeting.

Yveta Masarova, MS (Center Administrative Director) – effort on the project -
Yveta will manage all administrative aspects and financial budget preparation of the grant.

Sara Vassallo, BS (Research Technician, Church lab) effort on the project –
Sara is a senior technician. Her role involves providing technical support to post-doc and graduate students in the lab.

Francois Vigneault, PhD (Postdoctoral Fellow, Church lab), effort on the project – Salary support requested for in year 1. Francois has a fellowship (200702MFE- CIHR Fellowship award).

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): Church, George M.

Francois is a postdoctoral fellow who will be responsible for technology development related to single cell transcriptome analysis. More specifically his focus will be on developing library preparation protocols, in cell emulsion protocol, and in situ sequencing.

Dae Kim, PhD (Postdoctoral Fellow, Church lab) effort on the project - [EFFORT]

Dae Kim will focus on developing methods to achieve in-situ cell sequencing and single molecule in-cell sequencing. The methodology will encompass the integration of surface chemistries, attachment chemistries, and next-generation sequencing chemistries.

Jin Billy Li, PhD (Postdoctoral fellow, Church lab), effort on the project - [EFFORT]

Billy is a 4th year post-doc in the Church Lab. He will work on assessing allele specific expression in CTCHGV cell lines and on assessing causal mechanisms relevant to variations identified as causing allele specific expression such as differential RNA splicing, methylation, and degradation.

Je-Hyuk Lee, PhD (Postdoctoral Fellow, Church lab) effort on the project - [EFFORT]

Je-Hyuk Lee is a 2nd year post-doctoral fellow. He will be working on generating human induced Pluripotent Stem cells (iPS) and diverse cell types derived from them, and on analyzing them for allele-specific expression and other molecular traits.

Morten Sommer, PhD (Postdoctoral Fellow, Church lab) effort on the project - [EFFORT]

Morten Sommer is a first year post doc. Morten will be working on developing and applying technologies for characterizing human in situ genomics and transcriptomics. This will include development of novel approaches for high throughput gene identification and annotation.

Michael Sismour, PhD (Postdoctoral Fellow, Church lab) effort on the project - [EFFORT]

Michael Sismour. Bio-organic chemist. Development of methods for oligonucleotide synthesis, development and implementation of light labile chemistries, and development of conjugation chemistries for attachment of molecules/cells/beads to surfaces.

Tara Gianoulis, PhD (Postdoctoral Fellow, Church lab) effort on the project - [EFFORT]

Tara is a first year postdoc. She will develop and apply computational algorithms to CTCHGV data for inference of the causative status of DNA variations, for the analysis of large datasets (such as ribosome display datasets from Aim 4), and for design of large DNA constructs.

Daniel Levner, PhD (Engineer- Postdoctoral Fellow, Church lab) effort on the project - [EFFORT]

Dr. Levner holds a PhD in Electrical Engineering and specializes in optical systems and microfluidics. He will be involved in the development and implementation of integrated DNA sequencing and synthesis technology, and on cell handling technology, with focus on enhancement and development of instrumentation, as part of Aim 4 of the proposal.

Sasha Wait, PhD (Postdoctoral Fellow, Church lab) effort on the project - [EFFORT]

Sasha is a first year postdoctoral fellow. He will work on systems architecture and software development for data processing with a focus on data integration and data management. Sasha will help to improve variation discovery and analysis. He will also focus on interpretation of variants by combining data from numerous public and, where possible, private databases.

Graduate students:

Abraham Rosenbaum (graduate student, HMS) effort on the project - [EFFORT]

Budget Justification - HARVARD

Principal Investigator/Program Director (Last, first, middle): Church, George M.

Abraham is a 6th year grad student. Abraham will be working on in situ enzyme reactions, particularly ligation and rolling circle amplification, with the goal of in situ genomic and transcriptomic sequencing. He will also be collaborating with Dae Kim and Francois Vigneault to introduce barcodes into individual cells for single cell sequencing.

Madeleine Ball (graduate student, HMS) effort on the project -

Madeleine is currently a 7th year graduate student at Harvard Medical School. Madeleine's role on the project will be studying epigenetic aspects in cells and how these relate to our study of causal cis variants affecting gene expression. Madeleine will use high throughput targeted and genome-scale techniques that utilize next generation sequencing technology.

Xavier Rios: (graduate student, HMS) effort on the project -

Xavier is a 3rd year graduate student. Xavier will be working on developing MAGE-human techniques of oligonucleotide-mediated genome alterations which will be used for generating combinatorially altered cis regulatory regions on human cells. In addition, he will be developing on improvements of the Zinc-finger nuclease-mediated recombination.

Uri Laserson: (graduate student, MIT) effort on the project -

Uri is a 4th year graduate student at MIT. He will be working with Francois Vigneault on development of experimental protocols and computational methods for single cell analysis

Consultants:

Robert Green, PhD. (consultant)

Robert C. Green, MD, MPH, Dr. Green is currently professor of neurology, epidemiology and genetics at Boston University School of Medicine. Dr. Green will contribute to the CEGS by assisting in the training and mentoring of research fellows and advising the study on issues related to human studies that may arise.

Jeantine Lunshof, PhD (consultant)

Role on the project: Jeantine E. Lunshof, PhD, philosopher and bioethicist, will contribute to the analysis of the conceptual issues raised by fundamental research using the human as a model organism and the translation into personalized medicine. Besides, she will be involved with ethics consultancy concerning human subjects research, protocol development, and the innovation of procedures and content in ethics.

Jason Bobe (consultant)

Jason Bobe, CEGS Science Communication Coordinator. Jason will be responsible for the coordination of research activities and will serve as a liaison between the Church Lab, collaborators, and affiliates.

Jason Morrison (consultant)

Jason will work with Jason Bobe to develop software, including web applications, to support CEGS data management and dissemination.

JHV Consulting (Ward Vandewege) (consultant)

Role on the project: Specify, Maintain and setup cluster hardware and network configurations for Church lab 96 node data-intensive, high performance computing clusters. Develop security policies.

Tom Clegg (consultant)

Development and maintenance of software infrastructure for storage and processing of CEGS DNA sequence, transcriptome data and other data; development and maintenance of analysis tools for researchers; system administration for research systems and publicly accessible web services; technical support for cloud computing applications.

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): Church, George M.

Specialized Equipment: \$500,000/ yr 1

ITEM	estimate
Digital Micromirror Device (DMD, 1960x1080, UV version)	\$ 16,000
Total Internal Reflection Fluorescence (TIRF) microscope objective	\$ 15,000
Parts to attach DMD to Polonator	\$ 3,000
Microfluidics (10 devices and 2 chips)	\$ 13,000
Stem cells; BioLevigator benchtop bioreactor (cell handling)	\$ 16,000
Stem cells: Luminescence/Flourescence plate reader: (selection and testing of functional recombinants)	\$ 12,000
Stem cells Automated cell counter: (cell handling)	\$ 10,000
Stem cells: Inverted phase contrast & fluorescence microscope with video output for cell imaging)	\$ 16,000
Stem cells: Single chamber Nucleofector	\$ 12,000
Stem cells: Liquid nitrogen storage tank	\$ 5,000
Stem cells: fluidics device parts	\$ 10,000
automation for MAGE: 96 well nucleofector	\$ 60,000
automation for MAGE: plate reader	\$ 40,000
HPLC for chemistry for integrated DNA seq/synth + cell sorting	\$ 45,000
10 high performance compute nodes for CTCHGV to be added to Harvard cluster (based on specs: IBM System x3550, Dual core/Dual CPU Intel Xeon EM64T CPUs (4 cores) , 8 GB RAM, 2.73 GB SAS disks	\$ 60,000
New Polonator (for adaptation to synthetic biology and cell sorting: Aims 4.1, 4.3)	\$ 167,000

Supplies:**General lab supplies:**

Disposable supplies: In Year 1, the Church lab will support, not including George Church, approximately 9 postdocs, 3 staff, 4 grad students and coordinator. Each full-time researcher is expected to consume, on average \$19,000 worth of supplies per year * (projected itemization below).

Chemicals	\$ 1,500
Enzymes	4,000
Oligonucleotides	4,000
Sequencing charges	2,500
Cell sorting charges	1,000
Tissue culture supplies	1,000
Film	500
Plasticware	3,000
Glassware	500
Radioisotopes	<u>1,000</u>
	19,000 * ----- FTE

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): Church, George M.

Computers:

computer resources for investigators and lab personnel to support research activities, including software, licenses, fees.

Travel:

Domestic meetings: Funds are requested so that researchers may attend conference each year. The meetings will play an essential role in the training of the postdoctoral fellows and graduate students, especially given the interdisciplinary nature of the proposed projects. We estimate that the cost of a conference will be on average \$2,000 per person in Year 1 and will cover travel, housing, registration, meals, and incidentals.

Other:**Harvard Tuition: Graduate students:**

A Rosenbaum, M Price (2x5,070)	\$10,140
Xavier Rios (11,600)	\$11,600

HST Program (The Harvard-MIT Division of Health Sciences and Technology (HST) –Program of Massachusetts Institute of Technology (MIT), Harvard Medical School (HMS), Harvard University. Uri Laserson is a 4th year graduate student in HST program with advisor Dr. George Church.

Uri Laserson tuition & stipend	<div style="border: 1px solid black; padding: 2px;">Institutional Basic Salary</div>	\$43,896
Grad program fee (4x3,000)		\$12,000

Publications:	\$10,000
Service contracts:	\$36,000
Glass washing & media services:	\$5,000
Sequencing services:	\$215,405

Indirect costs /year 1:

HMS F&A Rate: 69.5%

Minority Initiatives:

Direct cost:	\$300,000
Indirect cost:	\$208,250

Budget - MAP-yr1

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY						FROM 4/1/10	THROUGH 3/31/11	
PERSONNEL (Applicant organization only)		Months Devoted to Project			INST.BASE SALARY	DOLLAR AMOUNT REQUESTED (omit cents)		
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths		SALARY REQUESTED	FRINGE BENEFITS	TOTAL
Church, George M.	PD/PI					0		0
Adeyemi Adesokan	PostDoc	EFFORT			Institutional Base Salary	41,796	11,285	53,081
TBH -PostDoc	PostDoc					40,500	10,935	51,435
Yveta Masarova	MAPS coordinator					32,789	13,771	46,560
TBH-PostBac	PostBac					31,200	8,424	39,624
TBN	student					4,000		4,000
TBN	student					4,000		4,000
TBN	student					4,000		4,000
TBN	student					4,000		4,000
TBN	student					4,000		4,000
TBN	student					4,000		4,000
SUBTOTALS →						170,285	44,415	214,700
CONSULTANT COSTS								
EQUIPMENT (Itemize)								
SUPPLIES (Itemize by category)								
Lab supplies: \$ 65,300								
Computers and software : \$ 10,000								
75,300								
TRAVEL								
Travel to a conference: 5 person x \$2,000								
10,000								
PATIENT CARE COSTS		INPATIENT						
		OUTPATIENT						
ALTERATIONS AND RENOVATIONS (Itemize by category)								
OTHER EXPENSES (Itemize by category)								
CONSORTIUM/CONTRACTUAL COSTS						DIRECT COSTS		
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD (Item 7a, Face Page)						\$ 300,000		
CONSORTIUM/CONTRACTUAL COSTS						FACILITIES AND ADMINISTRATIVE COSTS		
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD						\$ 300,000		

MAP - 5yr budget

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

**BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY**

BUDGET CATEGORY TOTALS		INITIAL BUDGET PERIOD (from Form Page 4)	ADDITIONAL YEARS OF SUPPORT REQUESTED			
			2nd	3rd	4th	5th
PERSONNEL: <i>Salary and fringe benefits. Applicant organization only.</i>		214,700	220,421	226,314	232,383	238,635
CONSULTANT COSTS						
EQUIPMENT						
SUPPLIES		75,300	77,559	79,886	82,282	84,751
TRAVEL		10,000	10,300	10,609	10,927	11,255
PATIENT CARE COSTS	INPATIENT					
	OUTPATIENT					
ALTERATIONS AND RENOVATIONS						
OTHER EXPENSES						
CONSORTIUM/ CONTRACTUAL COSTS	DIRECT					
SUBTOTAL DIRECT COSTS (Sum = Item 8a, Face Page)		300,000	308,280	316,808	325,593	334,640
CONSORTIUM/ CONTRACTUAL COSTS	F&A					
TOTAL DIRECT COSTS		300,000	308,280	316,808	325,593	334,640
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD						\$ 1,585,322

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Budget justification - MAP

Principal Investigator/Program Director (Last, first, middle): Church, George M.

MAP Budget Justification

Undergraduate Training program

EFFORT / year (EFFORT summer stipend x 6 students per year)

For summer 2010, we are planning to enroll 6 students as part of CEGS MAP. Based on the success of the program, we aim to continue with the program and hire 6 students per year.

Each student will receive a EFFORT summer stipend EFFORT

Postbac Training program

We aim to enroll one postbac student per year. EFFORT Salary requested including fringe benefits. The position is currently filled by Gerardo Gonzalez in the Church Lab.

Postdoc Training program

Adeyemi Adesokan - (PostDoctoral fellow) EFFORT

The Church Lab is currently supporting one postdoc- Adeyemi Adesokan. We aim to support two postdoctoral fellows per year. Several postdoctoral candidates are currently in advanced stages of the interview process.

Administrative Coordinator of Minority Initiatives

Yveta Masarova - (Administrative coordinator) EFFORT

Yveta is a Center Administrative Director in the Church lab. She has developed extensive experience in the administration of multi-institution research endeavors and academic environments. She is in continual contact with the principal investigator, investigators and key personnel of the center, program alumni, and administrators on other campuses. As the Administrative Coordinator of Minority Initiatives, Yveta will manage day-to-day administration of CEGS MAP activities, including responding by phone, email, and mail communications regarding our program; maintaining and updating center records related to MAP activities and alumni (for a minimum of 5 years); updating web site information; producing materials for recruitment activities and other program events; coordinating MAP activities at the institutions of center collaborators; and organizing the MAP evaluation meeting between the second and third years.

Travel

\$10,000 /year (\$2,000 / trip x 5 person) to cover the cost of traveling to conference

The center will send 1-2 representatives of the center to at least three events every year (1) New England Science Symposium (2) Biomedical Sciences Careers conference (3) biannual NIH MAP coordinator meetings.

Supplies

\$75,300/year

We are requesting support for research related supplies for 2 postdocs and 1 postbac and 6 undergraduate students; and also to support the activities of our coordinators, including general office supplies, printed materials, conference and event supplies.

Computers

\$10,000 / year (\$2000 x 5 computers/software licenses)

Direct cost: \$300,000

Indirect cost: \$208,250



Children's Hospital Boston

A teaching affiliate of Harvard Medical School

Office of Sponsored Programs
300 Longwood Avenue
Boston, Massachusetts 02115
phone 617-919-2729 | fax 617-730-0247
osp@childrens.harvard.edu
www.childrenshospital.org

STATEMENT OF INTENT

April 28, 2009

Investigator for Cooperating Institution: Dr. George Daley
Application Title: "Causal Transcriptional Consequences of Human Genetic Variation"
Period of Performance: 04/01/2010 – 03/31/2015
Total Costs: \$433,335 (\$250,000 DC; \$183,335 F&A)
Consortium Institution: Harvard Medical School
Investigator for Consortium Institution: Dr. George Church

To Whom It May Concern:

In signing below and offering to participate in this research program, the Cooperating Institution certifies that neither they nor their principals are presently debarred, suspended, proposed for debarment, declared ineligible or voluntarily excluded from receiving funds from any federal department or agency and are not delinquent on any federal debt; they are in compliance with the Drug Free Workplace Act of 1988; they are in compliance with U.S. Code, Section 1352, restrictions on the use of federal funds for the purpose of lobbying; they have filed annually with the Office of Scientific Integrity a PHS form 6349 governing Misconduct in Science; they have filed with DHHS compliance offices certification forms governing Civil Rights (441), Handicapped Individuals (641), Sex Discrimination (639-A), and Age Discrimination (680); they are in compliance with PHS policy governing Program Income; they have established policies in compliance with 45 CFR Part 46, Subpart A (protection of human subjects); the Animal Welfare Act (PL-89-544 as amended) and the Health Research Exchange Act of 1985 (Public Law 99-158); and that they are in compliance with NIH guidelines regarding human pluripotent stem cell research, transplantation of fetal tissue, recombinant DNA and human gene transfer research, and inclusion of women, children & minorities in research.

The appropriate programmatic and administrative personnel of each institution involved in this grant application are aware of the National Institutes of Health grant policies and are prepared to establish the necessary inter-institutional agreements consistent with those policies. In signing below, the Cooperating Institution certifies that it has implemented and is enforcing a written policy of Conflict of Interest consistent with the provisions of 42 CFR Part 50, Subpart F & 45 CFR Subtitle A, Part 94 and that there is no real or apparent conflict of interest for the purposes of this project.

Sincerely,

Signature

[Signature box]

Manager of Sponsored Programs & Senior Grant Officer

Form Approved Through 11/30/2010

OMB No. 0925-0001

Department of Health and Human Services Public Health Services Grant Application <i>Do not exceed character length restrictions indicated.</i>		LEAVE BLANK—FOR PHS USE ONLY. <table border="1" style="width:100%; border-collapse: collapse;"> <tr> <td style="width:33%;">Type</td> <td style="width:33%;">Activity</td> <td style="width:33%;">Number</td> </tr> <tr> <td colspan="2">Review Group</td> <td>Formerly</td> </tr> <tr> <td colspan="2">Council/Board (Month, Year)</td> <td>Date Received</td> </tr> </table>		Type	Activity	Number	Review Group		Formerly	Council/Board (Month, Year)		Date Received
Type	Activity	Number										
Review Group		Formerly										
Council/Board (Month, Year)		Date Received										
1. TITLE OF PROJECT (Do not exceed 81 characters, including spaces and punctuation.) Causal Transcriptional Consequences of Human Genetic Variation												
2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION <input type="checkbox"/> NO <input checked="" type="checkbox"/> YES (If "Yes," state number and title) Number: PAR-08-094 Title: Centers of Excellence in Genomics Science												
3. PROGRAM DIRECTOR/PRINCIPAL INVESTIGATOR		New Investigator <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes										
3a. NAME (Last, first, middle) Daley, George, Q.		3b. DEGREE(S) MD PhD										
3c. POSITION TITLE Associate Professor		3d. MAILING ADDRESS (Street, city, state, zip code) Children's Hospital Boston 300 Longwood Avenue Boston, MA 02115										
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Medicine		E-MAIL ADDRESS: george.daley@childrens.harvard.edu										
3f. MAJOR SUBDIVISION Hematology/Oncology												
3g. TELEPHONE AND FAX (Area code, number and extension) TEL: 617-919-2013 FAX: 617-730-0222												
4. HUMAN SUBJECTS RESEARCH <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes		4a. Research Exempt <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes If "Yes," Exemption No. E4										
4b. Federal-Wide Assurance No. FWA00002071		4c. Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes										
		4d. NIH-defined Phase III Clinical Trial <input type="checkbox"/> No <input type="checkbox"/> Yes										
5. VERTEBRATE ANIMALS <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		5a. Animal Welfare Assurance No. A3303-01										
6. DATES OF PROPOSED PERIOD OF SUPPORT (month, day, year—MM/DD/YY) From 04/01/10 Through 03/31/15		7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD 7a. Direct Costs (\$) \$47,089										
		7b. Total Costs (\$) \$80,757										
		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT 8a. Direct Costs (\$) \$250,000										
		8b. Total Costs (\$) \$433,335										
9. APPLICANT ORGANIZATION Name Children's Hospital Boston Address Children's Hospital Boston 300 Longwood Avenue Boston, MA 02115		10. TYPE OF ORGANIZATION Public: <input type="checkbox"/> Federal <input type="checkbox"/> State <input type="checkbox"/> Local Private: <input checked="" type="checkbox"/> Private Nonprofit For-profit: <input type="checkbox"/> General <input type="checkbox"/> Small Business <input type="checkbox"/> Woman-owned <input type="checkbox"/> Socially and Economically Disadvantaged										
		11. ENTITY IDENTIFICATION NUMBER 1042774441A1 DUNS NO. 076593722 Cong. District 8th										
12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE Name Theresa Applegate Title Manager of Sponsored Programs Address Children's Hospital Boston 300 Longwood Avenue Boston, MA 02115 Tel: 617-919-2729 FAX: 617-730-0247 E-Mail: osp @childrens.harvard.edu		13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION Name Theresa Applegate Title Manager of Sponsored Programs Address Children's Hospital Boston 300 Longwood Avenue Boston, MA 02115 Tel: 617-919-2729 FAX: 617-730-0247 E-Mail: osp @childrens.harvard.edu										
14. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.		SIGNATURE OF OFFICIAL NAMED IN 13. (In ink, "Pkg" signature not acceptable.) Signature _____										
		DATE 4/28/09										

Sub. Dr. George Daley

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY						FROM 4/1/10	THROUGH 3/31/11	
PERSONNEL (Applicant organization only)		Months Devoted to Project			INST.BASE SALARY Institutional Base Salary	DOLLAR AMOUNT REQUESTED (omit cents)		
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths		SALARY REQUESTED	FRINGE BENEFITS	TOTAL
George Daley	Principal Investigator	EFFORT				0	0	0
In-Hyun Park	Research Associate					0	0	0
TBN	Research Technician					17,500	5,250	22,750
SUBTOTALS →						17,500	5,250	22,750
CONSULTANT COSTS								
EQUIPMENT (Itemize)								
SUPPLIES (Itemize by category)								
Glassware, Plasticware, Disposables, Cell Culture supplies: \$9,000								
Chemicals and Reagents: \$10,339								
Computer Hardware and Software: \$5,000								
24,339								
TRAVEL								
PATIENT CARE COSTS		INPATIENT						
		OUTPATIENT						
ALTERATIONS AND RENOVATIONS (Itemize by category)								
OTHER EXPENSES (Itemize by category)								
CONSORTIUM/CONTRACTUAL COSTS						DIRECT COSTS		
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD (Item 7a, Face Page)						\$ 47,089		
CONSORTIUM/CONTRACTUAL COSTS						FACILITIES AND ADMINISTRATIVE COSTS		
						33,668		
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD						\$ 80,757		

Sub: Dr. George Daley

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

**BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY**

BUDGET CATEGORY TOTALS		INITIAL BUDGET PERIOD (from Form Page 4)	ADDITIONAL YEARS OF SUPPORT REQUESTED			
			2nd	3rd	4th	5th
PERSONNEL: <i>Salary and fringe benefits. Applicant organization only.</i>		22,750	23,432	24,135	24,859	25,605
CONSULTANT COSTS						
EQUIPMENT						
SUPPLIES		24,339	25,069	25,821	26,595	27,394
TRAVEL						
PATIENT CARE COSTS	INPATIENT					
	OUTPATIENT					
ALTERATIONS AND RENOVATIONS						
OTHER EXPENSES						
CONSORTIUM/ CONTRACTUAL COSTS	DIRECT					
SUBTOTAL DIRECT COSTS (Sum = Item 8a, Face Page)		47,089	48,501	49,956	51,455	52,999
CONSORTIUM/ CONTRACTUAL COSTS	F&A	33,668	35,405	36,967	38,076	39,219
TOTAL DIRECT COSTS		80,757	83,906	86,923	89,531	92,218
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD						\$ 250,000

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Personnel:

George Daley will serve as PI. He is providing EFFORT and taking no salary.In-Hyun Park will serve as a Research Associate. He is providing EFFORT and taking no salary. Private Source is the source of Dr. Daley's salary. Private Source is the source of Dr. Park's salary.

The Research Technician (TBN) will be skilled in the production of human induced pluripotent stem cells, which will be essential to providing immortal cells for each individual in the study. He or she will provide EFFORT and require \$22,750 in salary and fringe benefits, with a three percent increase annually.

Supplies:

\$9,000 for Glassware, Plasticware, Disposables, Cell Culture supplies will be used to grow cells/cell lines in culture.

\$10,339 for Chemicals and Reagents are used to prepare common reagents, monitor the growth of cells.

\$5,000 for Computer Hardware and Software will be necessary for the work of the research technician.

UNIVERSITY OF CALIFORNIA, SAN DIEGO

UCSD

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO



SANTA BARBARA • SANTA CRUZ

JAN MOLINA J.D.
CONTRACT NEGOTIATOR
TELEPHONE (858) 822-2901
FACSIMILE (858) 534-0280
E-MAIL: JMOLINA@UCSD.EDU

OFFICE OF CONTRACT AND GRANT ADMINISTRATION
University of California, San Diego
9500 Gilman Drive #0934
La Jolla, CA 92093-0934

OVERNIGHT MAIL DELIVERY ADDRESS:
10300 North Torrey Pines Road, Second Floor
La Jolla, CA 92037

CONSORTIUM STATEMENT OF INTENT

Research Proposal Entitled: Causal Transcriptional Consequences of Human Genetic Variation

Principal Investigator: Dr. Kun Zhang

Period of Performance: 4/01/10 – 3/31/15

Amount Requested: \$765,901

The appropriate programmatic and administrative personnel of each institution involved in this grant application are aware of the National Institutes of Health consortium grant policy and are prepared to establish the necessary inter-institutional agreement (s) consistent with that policy.

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Signature

[Signature box]

Name: Jan Molina
Title: Contract Negotiator
Date: 4.16.09

UCSD # 2009-3915

Form Approved Through 11/30/2010

OMB No. 0925-0001

Department of Health and Human Services Public Health Services Grant Application <i>Do not exceed character length restrictions indicated.</i>		LEAVE BLANK—FOR PHS USE ONLY. Type _____ Activity _____ Number _____ Review Group _____ Formerly _____ Council/Board (Month, Year) _____ Date Received _____	
1. TITLE OF PROJECT (Do not exceed 81 characters, including spaces and punctuation.) Causal Transcriptional Consequences of Human Genetic Variation			
2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION <input type="checkbox"/> NO <input checked="" type="checkbox"/> YES (If "Yes," state number and title) Number: PAR-08-094 Title: Centers of Excellence in Genomics Science			
3. PROGRAM DIRECTOR/PRINCIPAL INVESTIGATOR		New Investigator <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
3a. NAME (Last, first, middle) Zhang, Kun		3b. DEGREE(S) Ph.D.	
3c. POSITION TITLE Assistant Professor		3d. MAILING ADDRESS (Street, city, state, zip code) UCSD Bioengineering 9500 Gilman Dr MC 0412 La Jolla CA 92093-0412	
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Bioengineering		3f. MAJOR SUBDIVISION General Campus	
3g. TELEPHONE AND FAX (Area code, number and extension) TEL: 858-822-7876 FAX: 858-534-5722		E-MAIL ADDRESS: kzhang@ucsd.edu	
4. HUMAN SUBJECTS RESEARCH <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes		4a. Research Exempt If "Yes," Exemption No. <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
4b. Federal-Wide Assurance No. FWA00004495		4c. Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
4d. NIH-defined Phase III Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		5. VERTEBRATE ANIMALS <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
5a. Animal Welfare Assurance No. A3033-01		6. DATES OF PROPOSED PERIOD OF SUPPORT (month, day, year—MM/DD/YY) From 04/01/10 Through 03/31/15	
7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD 7a. Direct Costs (\$) 93,373		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT 7b. Total Costs (\$) 144,261 8a. Direct Costs (\$) 495,729 8b. Total Costs (\$) 765,901	
9. APPLICANT ORGANIZATION Name The Regents of the University of California Address University of California, San Diego - UCSD 9500 Gilman Drive, 0934 La Jolla, California 92093-0934		10. TYPE OF ORGANIZATION Public: → <input type="checkbox"/> Federal <input checked="" type="checkbox"/> State <input type="checkbox"/> Local Private: → <input type="checkbox"/> Private Nonprofit For-profit: → <input type="checkbox"/> General <input type="checkbox"/> Small Business <input type="checkbox"/> Woman-owned <input type="checkbox"/> Socially and Economically Disadvantaged	
11. ENTITY IDENTIFICATION NUMBER 1956006144A1 DUNS NO 80-435-5790 Cong. District 53		12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE Name Jan Molina Title Contract and Grant Officer Address UCSD / OCGA 9500 Gilman Drive, 0934 La Jolla, California 92093-0934 Tel: 858-822-2901 FAX: 858-534-0280 E-Mail: jmolina@ucsd.edu	
13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION Name Jan Molina Title Contract and Grant Officer Address UCSD / OCGA 9500 Gilman Drive, 0934 La Jolla, California 92093-0934 Tel: 858-822-2901 FAX: 858-534-0280 E-Mail: jmolina@ucsd.edu		14. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and I accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.	
SIGNATURE OF OFFICIAL NAMED IN 13. _____ DATE 4/22/09		PHS 398 (Rev. 11/07)	

Fac

23

Form Page 1

VIZ00099581

Sub. CCSD: Kun Zhang

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY						FROM 04/01/10	THROUGH 3/31/11		
PERSONNEL (Applicant organization only)		Months Devoted to Project			INST.BASE SALARY Institutional Base Salary	DOLLAR AMOUNT REQUESTED (omit cents)			
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths		SALARY REQUESTED	FRINGE BENEFITS	TOTAL	
Kun Zhang	PD/PI	EFFORT				10,192	1,702	11,894	
TBN	Postdoc					41,346	9,096	50,442	
SUBTOTALS →						51,538	10,798	62,336	
CONSULTANT COSTS									
EQUIPMENT (Itemize)									
SUPPLIES (Itemize by category) Materials, chemicals, computers & software									
								27,288	
TRAVEL								3,000	
PATIENT CARE COSTS		INPATIENT							
		OUTPATIENT							
ALTERATIONS AND RENOVATIONS (Itemize by category)									
OTHER EXPENSES (Itemize by category) NGN/Communication costs									
								749	
CONSORTIUM/CONTRACTUAL COSTS					DIRECT COSTS				
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD (Item 7a, Face Page)								\$ 93,373	
CONSORTIUM/CONTRACTUAL COSTS					FACILITIES AND ADMINISTRATIVE COSTS				50,888
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD								\$ 144,261	

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

**BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY**

BUDGET CATEGORY TOTALS		INITIAL BUDGET PERIOD (from Form Page 4)	ADDITIONAL YEARS OF SUPPORT REQUESTED				
			2nd	3rd	4th	5th	
PERSONNEL: <i>Salary and fringe benefits. Applicant organization only.</i>			51,538	53,602	55,748	57,977	60,295
CONSULTANT COSTS			10,798	11,230	11,680	12,148	12,633
EQUIPMENT							
SUPPLIES			27,288	27,573	27,842	28,096	28,333
TRAVEL			3,000	3,000	3,000	3,000	3,000
PATIENT CARE COSTS	INPATIENT						
	OUTPATIENT						
ALTERATIONS AND RENOVATIONS							
OTHER EXPENSES		749	769	789	810	831	
CONSORTIUM/ CONTRACTUAL COSTS	DIRECT						
SUBTOTAL DIRECT COSTS (Sum = Item 8a, Face Page)		93,373	96,174	99,059	102,031	105,092	
CONSORTIUM/ CONTRACTUAL COSTS	F&A	50,888	52,415	53,987	55,607	57,275	
TOTAL DIRECT COSTS		144,261	148,589	153,046	157,638	162,367	
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD						\$ 765,901	

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Sub_UCSD: Kun Zhang

Budget Justification (UCSD)**Personnel** **Total \$ 62,336/year****Kun Zhang, Ph.D.** (Principal Investigator, all years, salary requested) **\$ 11,894/year**

Dr. Kun Zhang will be PI of the UCSD research group. He will oversee the development and optimization methods for RNA allelotyping and full genome haplotyping.

TBN., (Postdoctoral associate) **\$50,442/year**

Chromosome preparation, design and fabricate microwell chips, in situ whole genome amplicon on single chromosomes, library construction and shotgun sequencing.

Reagents and supplies **Total \$27,288/year**

We request an annual budget of \$27,288 to cover enzymes, consumables, chemicals, software, user fees for the nanofabrication facility (Nano3) in UCSD, and publication cost.

Travel **\$3,000/year**

Domestic meetings: Funds are requested so that each researcher may attend a conference each year (total of 2 persons). We estimate that the cost of a conference will be, on average, \$1,500 per person and will cover travel, housing, registration, meals, and incidentals.

Other Expenses **\$749/year**

Funds of \$749/year are requested for communication costs for telephone and associated voice and data charges which are directly related to individuals working on the project.

STATEMENT OF INTENT

Prime PI: Church, George M. _____ Prime Institution: Harvard Medical School
 Project Title: Causal Transcriptional Consequences of Human Genetic Variation

Cooperating Institution: Massachusetts General Hospital
 PI Name: J. Keith Joung ERA Commons ID: ERA Commons User Name
 Tel: 617-726-9462 Fax: 617-726-5684
 Email: jjoung@partners.org
 Project Period: 04-01-2010 to 03-31-2015 Direct Costs: \$982240 F&A Costs: \$ 744775
 Performance site: MGH- 13th Street, Bldg 149 6th Floor, Charlestown MA 02129

Cooperating Institutional Business Contact Information:

Name: Caren Briggs
 Address: 101 Huntington Ave Suite 3
 Tel: 617-954-9302 Fax: 617-954-9850 Email: cbriggs1@partners.org

Project information:	Yes	No	Assurance#:	Approval Date or Pending:
Human subjects:	<input type="checkbox"/>	<input checked="" type="checkbox"/>	_____	_____
Vertebrate animals:	<input type="checkbox"/>	<input checked="" type="checkbox"/>	_____	_____
Human Embryonic Stem Cells:	<input type="checkbox"/>	<input checked="" type="checkbox"/>		
Inventions and Patents: (for renewal applications)	<input type="checkbox"/>	<input checked="" type="checkbox"/>		
Program Income:	<input type="checkbox"/>	<input checked="" type="checkbox"/>		

Certifications:

In signing below and offering to participate in this research program, the Cooperating Institution certifies that neither they nor their principals are presently debarred, suspended, proposed for debarment, declared ineligible or voluntarily excluded from receiving funds from any federal department or agency and are not delinquent on any federal debt; they are in compliance with the Drug Free Workplace Act of 1988; they are in compliance with U.S. Code, Section 1352, restrictions on the use of federal funds for the purpose of lobbying; they have filed annually with the Office of Scientific Integrity a PHS form 6349 governing Misconduct in Science; they have filed with DHHS compliance offices certification forms governing Civil Rights (441), Handicapped Individuals (641), Sex Discrimination (639-A), and Age Discrimination (680); they are in compliance with PHS policy governing Program Income; they have established policies in compliance with 45 CFR Part 46, Subpart A (protection of human subjects); the Animal Welfare Act (PL-89-544 as amended) and the Health Research Exchange Act of 1985 (Public Law 99-158); and that they are in compliance with NIH guidelines regarding human pluripotent stem cell research, transplantation of fetal tissue, recombinant DNA and human gene transfer research, and inclusion of women, children & minorities in research.

The appropriate programmatic and administrative personnel of each institution involved in this grant application are aware of the PHS-NIH consortium grant policies and are prepared to establish the necessary inter-institutional agreements consistent with those policies. In signing below, the Cooperating Institution certifies it has implemented and is enforcing a written policy of conflicts of interest consistent with the provisions of 42 CFR Part 50, Subpart F & 45 CFR Subtitle A, Part 94. If a conflict is identified by the Cooperating Institution during the period of the award contemplated under this agreement, the Cooperating Institution will report to the Prime Awardee the existence of the conflict, including the grant title, PI (if different from the investigator with the financial interest) and the specific method the Cooperating Institution adopts for addressing the conflict (managing, reducing, or eliminating it). The Cooperating Institution will rely on the Prime Awardee to report the existence of the conflict to NIH.

Cooperating Institution Business Official:

CAREN BRIGGS
GRANTS & CONTRACTS ADMINISTRATOR

Typed Name & Title

Date 4/29/09

Signature

Form Approved Through 11/30/2010

OMB No. 0925-0001

Department of Health and Human Services Public Health Services Grant Application <i>Do not exceed character length restrictions indicated.</i>		LEAVE BLANK—FOR PHS USE ONLY. <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 33%;">Type</td> <td style="width: 33%;">Activity</td> <td style="width: 34%;">Number</td> </tr> <tr> <td>Review Group</td> <td></td> <td>Formerly</td> </tr> <tr> <td>Council/Board (Month, Year)</td> <td></td> <td>Date Received</td> </tr> </table>		Type	Activity	Number	Review Group		Formerly	Council/Board (Month, Year)		Date Received
Type	Activity	Number										
Review Group		Formerly										
Council/Board (Month, Year)		Date Received										
1. TITLE OF PROJECT (Do not exceed 81 characters, including spaces and punctuation.) Causal Transcriptional Consequences of Human Genetic Variation												
2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION <input type="checkbox"/> NO <input checked="" type="checkbox"/> YES (If "Yes," state number and title) Number: PAR -08-094 Title: Centers of Excellence in Genomics Science												
3. PROGRAM DIRECTOR/PRINCIPAL INVESTIGATOR		New Investigator <input type="checkbox"/> No <input type="checkbox"/> Yes										
3a. NAME (Last, first, middle) Joung, Jae, Keith		3b. DEGREE(S) Ph.D MD										
3c. POSITION TITLE Associate Chief of Pathology		3h. eRA Commons User Name eRA Commons User Name										
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Pathology		3d. MAILING ADDRESS (Street, city, state, zip code) 13st Street Bldg 149 6th Floor Charlstown MA 02129										
3f. MAJOR SUBDIVISION												
3g. TELEPHONE AND FAX (Area code, number and extension) TEL: 617-726-9462 FAX: 617-726-5684		E-MAIL ADDRESS: jjoung@partners.org										
4. HUMAN SUBJECTS RESEARCH <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		4a. Research Exempt If "Yes," Exemption No. <input type="checkbox"/> No <input type="checkbox"/> Yes										
4b. Federal-Wide Assurance No. 00003136		4c. Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes										
		4d. NIH-defined Phase III Clinical Trial <input type="checkbox"/> No <input type="checkbox"/> Yes										
5. VERTEBRATE ANIMALS <input type="checkbox"/> No <input type="checkbox"/> Yes		5a. Animal Welfare Assurance No. A3596-01										
6. DATES OF PROPOSED PERIOD OF SUPPORT (month, day, year—MM/DD/YY) From 04-01-2010 Through 03-31-2015		7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD 7a. Direct Costs (\$) 192217										
		7b. Total Costs (\$) 328674										
		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT 8a. Direct Costs (\$) 982240										
		8b. Total Costs (\$) 1727014										
9. APPLICANT ORGANIZATION Name Massachusetts General Hospital Address (The General Hospital Corporation) 55 Fruit Street Boston, MA 02114-2696		10. TYPE OF ORGANIZATION Public: → <input type="checkbox"/> Federal <input type="checkbox"/> State <input type="checkbox"/> Local Private: → <input checked="" type="checkbox"/> Private Nonprofit For-profit: → <input type="checkbox"/> General <input type="checkbox"/> Small Business <input type="checkbox"/> Woman-owned <input type="checkbox"/> Socially and Economically Disadvantaged										
		11. ENTITY IDENTIFICATION NUMBER 1042697983A1 DUNS NO. 07-313-0411 Cong. District 009										
12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE Name F. Richard Bringham, M.D. Title Sr. Vice President, Medicine & Research Mgmt Address MGH Research Management 101 Huntington Avenue, Suite 300 Boston, MA 02199 Tel: (617) 954-9309 FAX: (617) 954-9850 E-Mail: MGH-G&C@partners.org		13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION Name Caren Briggs Title Grants and Contracts Administrator Address MGH Research Management 101 Huntington Avenue, Suite 300 Boston, MA 02199 Tel: (617) 954-9302 FAX: (617) 954-9850 E-Mail: Cgriggs1@partners.org										
14. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.		SIGNATURE OF OFFICIAL NAMED IN 13. (Print "Off" signature and representative) <div style="border: 1px solid black; width: 150px; height: 40px; margin-top: 10px;"></div>										
		DATE <div style="border: 1px solid black; width: 100px; height: 40px; margin-top: 10px; text-align: center;">5/1/09</div>										

SUB-MCH-Joung, J.K

Program Director/Principal Investigator (Last, First, Middle): CHURCH, George M.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY						FROM 04-01-10	THROUGH 03-31-11	
PERSONNEL (Applicant organization only)		Months Devoted to Project			INST.BASE SALARY Institutional Base Salary	DOLLAR AMOUNT REQUESTED (omit cents)		
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths		SALARY REQUESTED	FRINGE BENEFITS	TOTAL
Keith Joung	PD/PI	EFFORT				28,605	10,012	38,617
TBD	Postdoc					45,000	14,400	59,400
TBD	Research Tech					35,000	11,200	46,200
SUBTOTALS →						108,605	35,612	144,217
CONSULTANT COSTS								
EQUIPMENT (Itemize)								
96 WELL SHAKER								
15,000								
SUPPLIES (Itemize by category)								
33,000								
TRAVEL								
PATIENT CARE COSTS		INPATIENT						
		OUTPATIENT						
ALTERATIONS AND RENOVATIONS (Itemize by category)								
OTHER EXPENSES (Itemize by category)								
CONSORTIUM/CONTRACTUAL COSTS					DIRECT COSTS			
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD (Item 7a, Face Page)								\$ 192,217
CONSORTIUM/CONTRACTUAL COSTS					FACILITIES AND ADMINISTRATIVE COSTS			
								136,457
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD								\$ 328,674

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

**BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY**

BUDGET CATEGORY TOTALS		INITIAL BUDGET PERIOD (from Form Page 4)	ADDITIONAL YEARS OF SUPPORT REQUESTED			
			2nd	3rd	4th	5th
PERSONNEL: <i>Salary and fringe benefits. Applicant organization only.</i>		144,217	148,539	151,863	155,224	158,685
CONSULTANT COSTS						
EQUIPMENT		15,000				
SUPPLIES		33,000	42,000	43,260	44,558	45,894
TRAVEL						
PATIENT CARE COSTS	INPATIENT					
	OUTPATIENT					
ALTERATIONS AND RENOVATIONS						
OTHER EXPENSES						
CONSORTIUM/ CONTRACTUAL COSTS	DIRECT					
SUBTOTAL DIRECT COSTS (Sum = Item 8a, Face Page)		192,217	190,539	195,123	199,782	204,579
CONSORTIUM/ CONTRACTUAL COSTS	F&A	136,457	146,715	150,245	153,832	157,526
TOTAL DIRECT COSTS		328,674	337,254	345,367	353,614	362,106
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD						\$ 1,727,014

JUSTIFICATION: Follow the budget justification instructions exactly. Use continuation pages as needed.

Sub- MGH: Keith J. Joung

Principal Investigator/Program Director (Last, First, Middle): Church, George M.

Personnel:

J. Keith Joung, M.D., Ph.D., Principal Investigator EFFORT

Dr. Joung is Associate Chief of Pathology for Research, Director of the Molecular Pathology Unit, and a member of the Center for Cancer Research and the Center for Computational and Integrative Biology at Massachusetts General Hospital. He is also an Associate Professor of Pathology at Harvard Medical School. Dr. Joung has extensive training in studying protein-DNA interactions with a particular emphasis on understanding those mediated by Cys₂His₂ zinc fingers. Dr. Joung also has considerable experience in the engineering of synthetic zinc finger proteins: he is the inventor of the OPEN (for Oligomerized Pool ENgineering) zinc finger engineering technology and the bacterial two-hybrid selection system. In addition, Dr. Joung's lab has considerable experience using zinc finger nucleases to modify endogenous human genes in various cell types.

TBD, Post-doctoral Research Fellow EFFORT

A post-doctoral fellow with a strong molecular biology background will be recruited to conduct the experiments aimed at engineering additional OPEN zinc finger pools, at engineering a comprehensive set of zinc finger arrays, at improving the efficiency of zinc finger nuclease-enhanced homologous recombination (relative to undesired NHEJ events), at improving methods to delivery zinc finger nucleases to cells, and at developing the zinc finger nuclease-enhanced "segmental replacement" strategy. This individual will spend EFFORT of their time on these projects.

TBD, Research Technician EFFORT

A research technician, preferably with a strong molecular biology background, will be recruited to assist Dr. Joung and the post-doctoral fellow supported by this application to conduct the experiments outlined above. This individual will prepare specialized bacterial media and other required chemical reagents, construct bacterial strains, prepare competent cells, perform transformations and selections, isolate plasmid DNA, prepare samples for DNA sequencing, analyze sequences, transfect human cells in culture, perform genotyping assays, and track and enter data into a database. This individual will devote EFFORT of their time to this project.

Supplies (for Year 1)

Bacteriological medium components (\$1000): Components for standard LB and SOC medium. Components for making histidine-deficient NM medium required for OPEN selections including: amino acids, M9 salts, thiamine, magnesium chloride, zinc sulfate, calcium chloride, IPTG, and glucose. Also, IPTG for induction of zinc finger and alpha-gal4 hybrid protein expression in all media.

Enzymes for PCR and genotyping (\$7,000): Platinum PCR SuperMix Hi-Fidelity enzyme and various restriction enzymes needed for limited-cycle PCR/restriction digest assays and for limited-cycle PCR/DNA sequencing assays to assess mutation frequencies and to characterize specific genomic alterations.

Radioactive isotopes (\$1,000): ³²P-alpha-dNTPs for labeling DNA fragments by PCR in limited-cycle PCR-based assays.

Plasmid purification kits (\$8,000): Standard QIAgen mini prep, midi prep, and maxi prep kits for molecular biology.

Plasticware for bacteriologic and tissue culture work (\$10,000): Sterile serological pipets, sterile barrier pipet tips, sterile 96-well blocks, sterile 96-well plates, sterile 24-well blocks, bacterial Petri dish plates, sterile 10 cm x 10 cm plates, sterile Eppendorf tubes (1.5 ml and 2.0 ml), PCR tubes and 96-well PCR plates.

Sub- MGH: Keith J. Joung

Principal Investigator/Program Director (Last, First, Middle): Church, George M.

Transfection reagents for human cells (\$5,000): Nucleofection kits and Lipofectamine 2000 reagent required to introduce zinc finger nuclease-encoding and donor template DNAs into human cells.

Oligonucleotide synthesis (\$1,000): We will require oligonucleotides for selection experiments and construction of various expression plasmids.

Equipment (Year 1):

Microtitertron shaker for 96-well blocks/plates (\$15,000): Most of the zinc finger engineering and specialized molecular biology protocols developed in the Joung laboratory have been optimized for 96-well format. The work required for this application would exceed the capacity of a Microtitertron shaker that is already present and heavily used in the Joung lab. Therefore, we are requesting an additional machine that would be necessary to conduct the experiments of this proposal.

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

BIOGRAPHICAL SKETCHProvide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Church, George, M., PhD.		POSITION TITLE Professor	
eRA COMMONS USER NAME (credential, e.g., agency login) eRA Commons User Name			
EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)			
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY
Duke University, Durham, NC	B.A.	1974	Zoology & Chem.
Harvard University, Cambridge, MA	PhD.	1984	Biochem. & Mol. Biol.

A. Positions:

1984 Scientist, Biogen Research Corporation, Cambridge, MA
 1985-1986 Research Fellow, Anatomy, Univ. Calif., San Francisco, CA
 1986-1998 Assistant/Associate Professor of Genetics, Harvard Medical School, Boston, MA
 1997-present Director of the Lipper Center for Computational Genetics, Boston, MA
 1998-present Professor of Genetics, Harvard Medical School, Boston, MA
 2002-present Director of the Harvard/MIT DOE Genomes-to-Life Center
 2004-present Director of the Harvard/MIT/WashU NHGRI CEGS
 2006-present Senior Associate of Broad Inst. of Harvard & MIT (1990 Genome Center Co-founder)

Honors, Awards, & Scientific Memberships:

1974-1975 National Science Foundation Predoctoral Fellow
 1985-1986 Life Sciences Research Foundation Fellow
 1976 National Science Foundation Program Project Grant Review Committee
 1986-1997 Howard Hughes Medical Institute
 1988, 1992, 1994 Department of Energy Genome Project Grant Review Committee
 1990 NIH Genome Study Section Grant Review
 1990 Co-founder of MIT, Stanford, & GTC Genome Sequencing Centers
 1994-1997 National Center for Human Genome Research Review Committee
 2001-present NIH BISTI, Pioneer, DOE Sloan Fellow grant review committees
 Editorial Boards Nature/EMBO-MSB, Genome Biology, Omics, BioMedNet
 Editorial Reviewer Nature (&NG, NB), Science, PNAS, Genome Research, NAR
 Scientific Boards: LS9, 23andme, Knome, Genomatica, CodonDevices, Joule, CompleteGenomics
 2008 World Economic Forum Technology Pioneer Award (LS9 & 23andme)
 2009 American Society for Microbiology Biotechnology Research Award

B. Selected peer-reviewed publications. (also see <http://arep.med.harvard.edu>)

192. de Magalhaes JP, Curado J, Church GM (2009) Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*. Feb 2. PMID: 19189975
 178. Alexander Wait Zaranek, Tom Clegg, Ward Vandeweghe, George M. Church: Free Factories: Unified Infrastructure for Data Intensive Web Services. *USENIX Annual Technical Conference 2008*: 391-404
 177. de Magalhaes JP, Budovsky A, Lehmann G, Costa J, Li Y, Fraifeld V, Church GM (2009) Human Ageing Genomic Resources: online databases and tools for biogerontologists. *Aging Cell* 8(1):65-72.
 174. Schwartz D, Chou MF, Church GM (2008) Predicting protein post-translational modifications using meta-analysis of proteome-scale data sets. *Mol Cell Proteomics*. Oct 28.
 173. Lunshof JE, Chadwick R, Church GM (2008) Hippocrates revisited? Old ideals and new realities. *Genomic Med*. 2(1-2):1-3. PMID 18716915 - PMCID: PMC2518662

Program Director / Principal Investigator: Church, George M.

172. Snitkin ES, Dudley AM, Janse DM, Wong K, Church GM, Daniel Segre D (Sep 2008). Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. Genome Biology 9:R140.
171. Isenbarger TA, Finney M, Church GM, Gilbert W, Carr CE, Zuber MT, Ruvkun G. The most conserved genome segments for life detection on Earth and other planets. Orig Life Evol Biosph. 2008 Oct 14.
170. Forest CR, Rosenbaum AM, Church GM (2008) DNA Sequencing By Ligation On Surface-Bound Beads In A Microchannel Environment MicroTAS 2008
169. Vigneault F, Sismour AM, Church GM. (Sep 2008) Efficient microRNA capture and barcoding via enzymatic oligonucleotide adenylation. Nature Methods 5, 777 - 779. PMID: 18711363
168. Church GM, Porreca GJ, Terry RC, Lares M (2008) High-speed imaging for DNA sequencing. Biophotonics
167. Dantas G, Sommer MO, Oluwasegun RD, Church GM. Bacteria subsisting on antibiotics. Science. 2008 Apr 4;320(5872):100-3.
163. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. From genetic privacy to open consent. Nat Rev Genet. 2008
166. Bang D, Church GM. Gene synthesis by circular assembly amplification. Nat Methods. 2008 Jan;5(1):37-9.
165. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW. Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus. PLoS Genet. 2007 Dec 21;3(12):e231
164. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Yuan Gao Y, Church GM, Shendure, J (2007) Multiplex amplification of large sets of human exons. Nat Methods. 2007 Nov;4(11):931-6. (Supplement)
163. de Magalhaes JP, Sedivy JM, Finch CE, Austad SN, Church GM (2007). A Proposal to Sequence Genomes of Unique Interest for Research on Aging J Gerontol A Biol Sci Med Sci. 62: 583-584
162. Nardi V, Raz T, Chao X, Wu CJ, Stone RM, Cortes J, Deininger MWN, Church G, Zhu J and Daley GQ (2007) Monitoring resistance to kinase inhibitors with polony technology: towards personalized therapy for CML patients. Oncogene. 6 Aug issue. (Supplement)
161. Conrad C, Zhu J, Conrad C, Schoenfeld D, Fang Z, Ingelsson M, Stamm S, Church G, Hyman BT (2007) Single Molecule Profiling of tau Gene Expression in Alzheimer's Disease. J. Neurochem Aug 28 issue.
160. Wright MA, Kharchenko P, Church GM, Segre D. Chromosomal periodicity of evolutionarily conserved gene pairs. Proc Natl Acad Sci U S A. 2007 Jun 19;104(25):10559-64.
159. Debbie Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T, Kettler G, Sullivan MB, Steen R, Hess WR, Church GM, Sallie W. Chisholm SW (2007) Genome-Wide Expression Dynamics of a Marine Virus and its Host Reveal Features of Co-evolution. Nature. 449(7158):83-6. (Supplement).
158. Bugl H, Danner JP, Molinari RJ, Mulligan JT, Park H-O, Reichert B, Roth DA, Wagner R, Budowle B, Scripp RM, Smith JAL, Steele SJ, Church G, Endy D (2007) DNA synthesis and biological security Nat Biotechnol. 25(6):627-629.
157. Kim JB, Porreca GJ, Gorham JM, Church GM, Seidman CE, Seidman JG (2007) Polony multiplex analysis of gene expression (PMAGE) in a mouse model of hypertrophic cardiomyopathy. Science Jun 8; 316(5830):1481-4.
156. Bakal C, Aach J, Church GM, Perrimon, N (2007) Defining the Components Of Local Signaling Networks That Regulate Cell Morphology Using Quantitative Morphological Signatures. Science Jun 22; 316(5832):1753-6. Supplement
155. de Magalhaes JP, Church GM. (2007) Analyses of human-chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. Mech Ageing Dev. 128(5-6):355-64.
154. Magalhaes JP, Costa J, Church GM (2007) An analysis of the relation between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts, J Gerontol A Biol Sci Med Sci. 62(2):149-60.
147. Forster AC, Church GM. (2006) Synthetic biology projects in vitro. Genome Res. 17(1):1-6.
146. Reppas NB, Wade JT, Church GM, Struhl K. (2006) The Transition between Transcriptional Initiation and Elongation in E. coli Is Highly Variable and Often Rate Limiting. Molecular Cell 24(5):747-757
145. Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, Steen R, Church GM, Chisholm SW. Global gene expression of Prochlorococcus ecotypes in response to changes in nitrogen availability. Mol Syst Biol. 2006;2:53. Epub 2006 Oct 3.

Program Director / Principal Investigator: Church, George M.

144. Derti A, Roth FP, Church GM, Wu CT. (2006) Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet.* 38(10):1216-20.
 143. Turner DJ, Shendure J, Porreca G, Church G, Green P, Tyler-Smith C, Hurles ME (2006) Assaying chromosomal inversions by single-molecule haplotyping. *Nat Methods.* Jun;3(6):439-45.
 142. Jaffe, J, Mani, DR, Leptos, K, Church, GM, Carr SA (2006) PEPPer: A platform for experimental proteomic pattern recognition. *Mol Cell Proteomics.* 2006 Jul 19;
 141. Su-In Lee, S-I, Pe'er, D, Dudley, AM, Church, GM, Daphne Koller D (2006) Identifying Regulatory Mechanisms using Individual Variation Reveals Key Role for Chromatin Modification. *Proc Natl Acad Sci USA* 103(38):14062-7
 140. Yang A, Zhu Z, Kapranov P, McKeon F, Church GM, Gingeras TR, and Kevin Struhl K (2006) Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol Cell.* 2006 Nov 17;24(4):593-602.
 139. Kuo W, Liu F, Jenssen T, Trimarchi J, Punzo C, Lombardi M, Sarang J, Maysuria M, Serikawa K, Lee SY, McCrann D, Kang J, Shearstone J, Burke J, Park D, Choi S, Perrin S, Church GM, Bumgarner R, Cepko, C (2006) A sequence oriented comparison of gene expression measurements across different hybridization-based technologies. *Nature Biotech.* Jul;24(7):832-840.
 138. Zhang, K, Martiny, AC, Reppas, NB, Barry, KW, Malek, J, Chisholm, SW, Church, GM (2006) Sequencing genomes from single cells via polymerase clones. *Nature Biotech.* Jun;24(6):680-6.
 137. Forster, AC & Church, GM (2005) Toward Synthesis of a Minimal Cell. *Nature-EMBO-Molecular Systems Biology* 2 doi:10.1038/msb4100090. (supplement).
 136. Zhang, K, Zhu, J, Shendure, J, Porreca, GJ, Aach, JD, Mitra, RD, Church, GM (2006) Long-range polony haplotyping of individual human chromosome molecules. *Nature Genetics* Mar; 38(3):382-7.
 135. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics.* 2006 Mar 29;7(1):177
-
133. Estrada B, Choe SE, Gisselbrecht SS, Michaud S, Raj L, Busser BW, Halfon MS, Church GM, Michelson AM. (2006) An Integrated Strategy for Analyzing the Unique Developmental Programs of Different Myoblast Subtypes. *PLoS Genet.* 2(2):e16
 132. deMagalhaes, J, Church, GM (2006) Cells discover fire: employing reactive oxygen species in development and aging. *Experimental Gerontology* Oct 11.
 131. Leptos, KC, Sarracino, DA, Church, GM (2006) MapQuant: Open-Source Software for Large-Scale Protein Quantitation. *Proteomics* 6(6):1770-82. PMID: 16470651. Data and software supplement.
 130. Shendure, J, Porreca, GJ, Reppas, NB, Lin, X, McCutcheon, JP, Rosenbaum, AM, Wang, MD, Zhang, K, Mitra, RD, Church, GM (2005) Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome *Science* 309(5741):1728-32.
 129. Wade JT Reppas NB Church GM & Struhl K (2005) Genomic analysis of LexA binding reveals the permissive nature of the Escherichia coli genome and identifies unconventional target sites. *Genes Dev.* 2005 Nov 1;19(21):2619-30.
 128. Gao, Y & Church G (2005) Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* 21: 3970-5.
 127. Lindell, D, Jaffe, JD, Johnson, ZI, Church, GM & Chisholm, SW (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438:86-9.

C. Research Support:

Ongoing:

DE-FG02-03ER63445 (GTL)

2/01/2003 – 1/31/2011

DOE-GTL

PI: George Church Lab

Title: Microbial Ecology, Proteogenomics & Computational Optima

Our role in this Project is to study proteomics and cell models for *Prochlorococcus* and *Pseudomonas*

P50 HG003170 (CEGS)

5/01/2004 – 4/30/2009

Program Director / Principal Investigator: Church, George M.

NIH - NHGRI

PI: George Church

Title: Molecular and Genomic Imaging Center

Our role is to develop polony technology for resequencing of DNA and RNA in stem cells

SA5283-11210 (NSF)

7/01/2006 – 6/30/2011

NSF

PI: Jay Keasling (UC Berkeley)

Title: Synthetic Biology Engineering Research Center (SynBERC)

Our role is to develop synthetic bacterial genome "chasses" for safe use in mammals

5U24CA126554 (CGCC)

9/28/2006 – 8/31/2009

NIH - NIC

PI: Raju Kucherlapati (BWH)

Title: Cancer Genome Characterization Centers

Our role is to quantitate RNA for cancer by using polony sequencing

W911NF-08-1-0254 (DARPA)

6/27/2008 - 4/30/2010

PI: Neil Gershenfeld (MIT)

Title: " Milli-Biology: Programmed Assembly of Engineered Materials"

Private Source

Private Source

10/01/2008 - 9/30/2009

PI: George Church

Private Source

Goal: Goal is to identify, study and limit emerging drug resistance using a combined high-throughput experimental and computational strategy.

RO1 HL 094963- 01 (NHLBI)

NIH - NHLBI

9/30/2008 - 6/30/2010

PI: George Church

Title: Targeted 2nd generation sequencing in phenotyped Framingham & PGP populations

Goal: We propose to develop, demonstrate, and validate a pipeline for high-throughput, low-cost targeted resequencing of all human exons based on next-generation (gen2) sequencing techniques in support of the long term goal of enabling sequencing to be used routinely to characterize genotypes and genetic variation in genome-wide medical targets for large populations of individuals

Private Source

11/01/2008 -10/30/2012

PI: George Church

Private Source

The main goal of this project is the identification and characterization of naked mole-rat genes that contributed to the evolution of a long lifespan in this species.

OVERLAP: None

Completed:

Private Source

1/01/2007 – 12/31/2007

PI: George Church

Private Source

Our role is sequencing personal genomes with new technology

Program Director/Principal Investigator (Last, First, Middle): Church, George

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME George Q. Daley		POSITION TITLE Samuel E. Lux, IV Chair in Hematology Director, Stem Cell Transplantation Program, Children's	
eRA COMMONS USER NAME	eRA Commons User Name		
EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, and include postdoctoral training.)			
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY
Harvard University	A.B.	1982	Biology
Massachusetts Institute of Technology	Ph.D.	1989	Biology
Harvard Medical School	M.D.	1991	Medicine

Professional Experience

1995-2003 Whitehead Fellow (Research), Whitehead Institute, Cambridge, MA.
Assistant Professor of Medicine, Harvard Medical School
2004 Associate Professor, Biological Chemistry, Pediatrics Harvard Medical School
2009 Director, Stem Cell Transplantation Program, Children's Hospital Boston
Samuel E. Lux, IV Chair in Hematology

Honors and Awards

1991 *summa cum laude*, Harvard Medical School
Leon Reznick Memorial Research Prize, Harvard Medical School
1992 National Institutes of Health Research Award for Clinical Trainees
New England Cancer Society Research Award
1996 Burroughs Wellcome Fund Career Award in the Biomedical Sciences
1997 Edward Mallinckrodt, Jr. Foundation Scholar Award
1999 Leukemia and Lymphoma Society of America Scholar Award
2003 Elected to American Society of Clinical Investigation
Burroughs Wellcome Clinical Scientist Award in Translational Research
2004 Stohman Scholar Award, Leukemia and Lymphoma Society
NIH Director's Pioneer Award
2006 Daland Prize, American Philosophical Society (achievement in patient-oriented research)
Elected Fellow of the American Association for the Advancement of Science
2007-2008 President, International Society for Stem Cell Research
2008 Investigator, Howard Hughes Medical Institute
Mentoring Award, Harvard Division of Biological and Biomedical Sciences
2009 E. Mead Johnson Award, Pediatric Research Society (for contributions in stem cell biology)
Member, Development 2 (DEV2) study section, NIH

Publications (selected from 194)

•Ben-Neriah Y, Bernards A, Paskind M, Daley GQ, Baltimore D. Alternative 5' exons in c-abl mRNA. **Cell** 1986; 44:577-586.
•Ben-Neriah Y, Daley GQ, Mes-Masson A-M, Witte ON, Baltimore D. The chronic myelogenous leukemia-specific P210 protein is the product of the bcr/abl hybrid gene. **Science** 1986; 233:212-214.
•Daley GQ, McLaughlin J, Witte ON, Baltimore D. The CML-specific P210 BCR/ABL protein, unlike v-abl, does not transform NIH/3T3 fibroblasts. **Science** 1987; 237:532-535.
•Daley GQ, Baltimore D. Transformation of an interleukin 3-dependent hematopoietic cell line by the chronic myelogenous leukemia-specific P210 BCR/ABL protein. **PNAS** 1988; 85:9312-9316.
•Daley GQ, Van Etten RA, Baltimore D. Induction of chronic myelogenous leukemia in mice by the P210 BCR/ABL gene of the Philadelphia chromosome. **Science** 1990; 247:824-830.
•Daley GQ, Van Etten RA, Baltimore D. Blast crisis in a murine model of chronic myelogenous leukemia. **PNAS** 1991; 88:11335-11338.
•Klucher K, Lopez D, Daley GQ. Secondary mutation maintains the transformed state in BaF3 cells with inducible BCR/ABL expression. **Blood** 1998; 91: 3927-3934.

Program Director/Principal Investigator (Last, First, Middle): Church, George

- Ghaffari S, Wu H, Gerlach MJ, Han Y, Lodish H, Daley GQ. BCR-ABL and v-SRC tyrosine kinase oncoproteins support normal erythroid development in Erythropoietin Receptor-deficient cells. **PNAS** 1999 96:13186-13190.
- Peters DG, Klucher KM, Perlingeiro RCR, Dessain SK, Koh EY, Daley GQ. Autocrine and paracrine effects of an ES-cell derived, BCR/ABL-transformed hematopoietic cell line that induces leukemia in mice. **Oncogene** 2001 20:2636-46.
- Perlingeiro RCR, Kyba M, Daley GQ. Clonal analysis during embryoid body differentiation reveals a common progenitor with primitive erythroid and adult lymphoid-myeloid hematopoietic potential. **Development** 2001 128:2641-4.
- Rideout WB, Hochedlinger K, Kyba M, Daley GQ, Jaenisch R. Correction of a genetic defect by nuclear transplantation and combined cell and gene therapy. **Cell** 2002 109:17-27.
- Kyba M, Perlingeiro RCR, Daley GQ. HoxB4 confers definitive lymphoid-myeloid engraftment potential on embryonic stem cells and yolk sac hematopoietic progenitors. **Cell** 2002 109:29-37.
- Azam M, Latek RR, and Daley GQ. Mechanisms of autoinhibition and STI571/Imatinib resistance revealed by mutagenesis of BCR/ABL. **Cell** 2003 112: 831-43.
- Perlingeiro RCR, Kyba M, Bodie S, Daley GQ. A role for thrombopoietin in hemangioblast development. **Stem Cells** 2003, 21: 272-80.
- Kyba M, Perlingeiro RCR, Hoover RR, Lu C-W, Pierce J, Daley GQ. Enhanced hematopoietic differentiation of ES cells conditionally expressing Stat5. **PNAS** 2003 100 (supp 1): 11904-10.
- Davidson AJ, Ernst P, Wang Y, Dekens MPS, Kingsley PD, Palis J, Korsmeyer SJ, Daley GQ, Zon LI. Cdx4 mutants fail to specify blood progenitors and can be rescued by multiple hox genes. **Nature** 2003 425:300-306.
- Geijsen N, Horoschak M, Kim K, Gribnau J, Eggan K, Daley GQ. Derivation of embryonic germ cells and male gametes from embryonic stem cells. **Nature** 2004 427: 148-154.
- Prudhomme W, Daley GQ, Zandstra P, Lauffenburger DA. Multivariate proteomic analysis of murine embryonic stem cell self-renewal versus differentiation signaling. **PNAS** 2004 101: 2900-2905.
- Daheron L, Opitz SI, Zaehres H, Lensch MW, Itskovitz-Eldor J, Daley GQ. LIF-STAT3 signaling is insufficient to maintain self-renewal of human embryonic stem cells. **Stem Cells** 2004, 22(5):770-8.
- Woolf PJ, Prudhomme W, Daheron L, Daley GQ, Lauffenburger DA. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. **Bioinformatics** 2004 21:741-53.
- Zaehres H, Lensch MW, Daheron L, Stewart SA, Itskovitz-Eldor J, Daley GQ. High efficiency RNA interference in human embryonic stem cells. **Stem Cells** 2005, 23:299-305.
- Wang Y, Yates F, Naveiras O, Ernst P, Daley GQ. Embryonic stem cell derived hematopoietic stem cells. **PNAS** 2005 102(52):19081-19086.
- Azam M, Nardi V, Shakespeare WC, Latek RR, Bohacek RS, Veach DR, Bornmann W, Clarkson B, Sawyer TK, Daley GQ. Activity of dual SRC-ABL inhibitors highlights role of BCR/ABL kinase dynamics in imatinib resistance **PNAS** 2006 103: 9244-9249; Jun 5; [Epub ahead of print].
- Kim K, Lerou P, Yabuuchi A, Lengerke C, Ng K, West J, Kirby A, Daly M, Daley GQ. Histocompatible parthenogenetic murine embryonic stem cells. **Science** 2007 Jan 26;315(5811):482-6.
- Daley GQ, et al. The ISSCR guidelines for human embryonic stem cell research **Science**. 2007 315:603-4.
- North TE, Goessling W, Walkley CR, Lengerke C, Kopani KR, Lord AM, Weber G, Bowman TV, Jang IH, Grosser T, Fitzgerald GA, Daley GQ, Orkin SH, Zon LI. Prostaglandin E2 regulates vertebrate hematopoietic stem cell homeostasis. **Nature** 2007 447(7147):1007-11.
- Daley GQ Gametes from embryonic stem cells: is the cup half empty or half full? **Science** 2007 316:409-10.
- Kim K, Ng K, Rugg-Gunn PJ, Shieh J-H, Kirak O, Jaenisch R, Wakayama T, Moore MA, Pederson RA, Daley GQ. Recombination signatures distinguish embryonic stem cells derived by parthenogenesis and somatic cell nuclear transfer. **Cell Stem Cell** 2007 1: 346-35.
- Lengerke C, Schmitt S, McKinney-Freeman S, Bowman TV, Davidson A, Green J, Zon LI, Daley GQ. BMP and Wnt specify mesoderm to hematopoietic fate by activation of the cdx-hox pathway **Cell Stem Cell** 2008 2: 72-82. Accepted prior to April 7, 2008
- Park I-H, Zhao R, West JA, Yabuuchi A, Huo H, Ince TA, Lerou PH, Lensch MW, and Daley GQ. Reprogramming of human somatic cells to pluripotency with defined factors. **Nature** 2008 451(7175):141-6. Epub 2007 Dec 23.
- Lerou PH, Yabuuchi A, Shea J, Takeuchi A, Cimini T, Ince T, Ginsburg E, Racowsky C, Daley GQ. Human embryonic stem cell derivation using poor quality embryos. **Nature Biotechnology** 2008 26:212-4.

Program Director/Principal Investigator (Last, First, Middle): Church, George

- *Chan EM, *Yates F, Boyer LF, Schlaeger TM, Daley GQ. Enhanced plating efficiency of human embryonic stem cells due to trypsin adaptation is reversible and independent of trisomy 12/17. **Cloning and Stem Cells** 2008 10(1):107-18. Accepted prior to April 7, 2008
- *Viswanathan S, *Daley GQ, *Gregory RI. Selective blockade of microRNA processing by Lin-28. **Science** 2008 4;320(5872):97-100 *co-corresponding authors. Accepted prior to April 7, 2008
- *Daley GQ, Scadden DT. Prospects for stem-cell based therapies. **Cell** 2008 132:544-548. Accepted prior to April 7, 2008
- *McKinney-Freeman SL, *Lengerke C, Jang IH, Schmitt S, Wang Y, Philitas M, Daley GQ. Modulation of murine embryonic stem cell-derived CD41+c-kit+ hematopoietic progenitors by ectopic expression of Cdx genes. **Blood** 2008 111(10): 4944-53. PMCID: PMC2384126 [Available on 5/15/2009]
- *Lerou PH, Yabuuchi A, Huo H, Miller JD, Boyer LF, Schlaeger TM, Daley GQ. Derivation and maintenance of human embryonic stem cells from poor-quality in vitro fertilization embryos. **Nature Protocols** 2008 3(5): 923-33. PMID: 18451800
- *Wang Y, Yabuuchi A, McKinney-Freeman S, Ducharme DMK, Scott GJ, Ward T, Ray M, Chawengsaksophak K, Archer TK, Daley GQ. Cdx deficiency compromises embryonic hematopoiesis in mouse. **PNAS (USA)** 2008 105(22):7756-61. PMCID: PMC2409377 [Available on 12/3/2008]
- *Lu C-W, Yabuuchi A, Chen L, Viswanathan S, Kim K, Daley GQ. Ras-Mitogen Activated Protein Kinase Signaling Promotes Trophoblast Formation from Embryonic Stem Cells and Murine Embryos. **Nature Genetics** 2008 40:921-6. PMID: 18536715
- *Azam M, Seeliger MA, Gray N, Kuriyan J, Daley GQ. Kinase activation by mutation of the gatekeeper threonine. **Nature Structural and Molecular Biology** 2008 15(10):1109-18. PMID: 18794843
- *Park I-H, Lerou PH, Zhao R, Daley GQ. Generation of human induced pluripotent stem cells. **Nature Protocols** 2008 3(7):1180-1186. PMID: 18600223
- *Park I-H, Arora N, Huo H, Maherali N, Ahfeldt T, Shimamura A, Lensch MW, Cowan C, Hochedlinger K, Daley GQ. Disease-specific induced pluripotent stem cells. **Cell** 2008 134(5):877-86. PMID: 18691744
- *Nardi V, Naveiras O, Azam M, Villuendas R, Castagnetti F, Martinelli G, and Daley GQ. Interferon mediated immune protection against Bcr-Abl induced leukemia requires the CCL6 and CCL9 chemokines. **Blood** 2009; January 26th online.
- *Daley GQ. Common themes of dedifferentiation in somatic cell reprogramming and cancer. **Cold Spring Harbor Symposium on Quantitative Biology** 2009; Jan 15 (online ahead of print).
- *Loh Y-H, Agarwal S, Park I-H, Urbach A, Huo H, Heffner GC, Miller JD, Daley GQ. Generation of induced pluripotent stem cells from human blood. **Blood** 2009 (March 18th online).
- *Deng J, Shoemaker R, Xie B, Gore A, Leproux E, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y, Zhang K. Digital DNA methylation analysis of stem cell reprogramming by targeted bisulfite sequencing. **Nature Biotechnology** 2009 27(4):353-60.
- *Ball MP, Li JB, Gao Y, Lee J, LeProust E, Park I-H, Xie B, Daley GQ, Church GM. Targeted and whole-genome methylomics reveals gene-body signatures in human cell lines. **Nature Biotechnology** 2009 27(4):361-8.
- *Goessling W, North TE, Loewer-Schmitt S, Lord AM, Lee S, Stoick-Cooper C, Puder M, Daley GQ, Moon RT, Zon LI. Genetic interaction between PGE2 and wnt/b-catenin signaling regulates developmental specification of stem cells and regeneration. **Cell** 2009 136: 1136-47.
- *Viswanathan S, Powers JT, Einhorn W, Hoshida Y, Toffanin S, Mermal CH, Lu J, Shah SP, Beroukhi R, Tanwar PS, Azam M, Perez-Atayde A, Teixeira J, Meyerson M, Frazier AL, Mullighan CG, Llovet JM, Radich J, Golub TR, Sorensen P, Daley GQ. Lin28 Enhances Tumorigenesis and is Associated With Advanced Human Malignancies. **Nature Genetics** (in press).
- *Baek K-H, Zaslavsky A, Lynch RC, Britt C, Okada Y, Siarey RJ, Lensch MW, Park I-H, Yoon SS, Minami T, Reeves R, Korenberg JR, Folkman J, Daley GQ, Aird WC, Galdzicki Z, Ryeom S. Down syndrome suppression of tumor growth and the role of the calcineurin inhibitor DSCR1. **Nature** (in press).
- *Naveiras O, Nardi V, Wenzel PL, Fahey F, Daley GQ. Bone marrow adipocytes as negative regulators of the hematopoietic microenvironment. **Nature** (in press).
- *Adamo L, *Naveiras O, McKinney-Freeman S, Mack PJ, Suchy-Dacey A, Yoshimoto M, Lensch MW, Yoder MC, #Garcia-Cardena G, #Daley GQ. Biomechanical forces enhance embryonic hematopoiesis. **Nature** (in press). *equal contribution/ #co-corresponding authors.

Program Director/Principal Investigator (Last, First, Middle): Church, George

C. Research Support
Ongoing Research Support

NIH R01 DK70055 Role: PI 5/15/05-4/30/10

Murine Models for Regenerative Medicine

This grant aims to develop murine knock-out models to explore the role of morphogens and homeodomain proteins in regulating hematopoiesis in both embryonic and adult contexts, and to use nuclear transfer and parthenogenesis to model treatment of genetic disorders of the blood and bone marrow.

NIH R01 DK59279 Role: PI 8/1/05-5/31/10

Derivation of Hematopoietic Stem Cells From Totipotent Embryonic Stem Cells

This grant aims to define the phenotypic and functional relationships between hematopoietic stem cells (HSCs) derived from embryos and embryonic stem cells (ESCs).

NIH R01 OD00256 Role: PI 9/30/04-7/31/09

NIH Director's Pioneer Award

The Pioneer Award supports studies of embryonic and germ cell fate specification, epigenetic reprogramming, and somatic cell reprogramming.

Private Source Role: PI 10/1/07-9/30/11

Private Source

This translational research award supports studies of drug resistance in chronic myeloid leukemia patients and studies of mechanisms of disease progression.

Private Source Role: PI 1/1/04-12/31/09 (No cost extension)

This is a career development award that supports translational studies in leukemia and stem cell biology.

Private Source co-PI 2/15/2009-2/14/2010

Private Source

This pilot award provides funding for derivation of iPS cells from up to five patients with immune deficiency.

Private Source 2/08-9/13

Private Source

Completed Research Support

R01 HL71265 NIH 8/1/02-7/31/06

Modeling Developmental Hematopoiesis With Embryonic Stem Cells

Role: PI

R01 CA86991 NIH 3/30/01-3/31/08

Therapeutic Mechanisms in Chronic Myelogenous Leukemia

Role: PI

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

BIOGRAPHICAL SKETCHProvide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Kun Zhang		POSITION TITLE Assistant Professor of Bioengineering	
eRA COMMONS USER NAME (credential, e.g., agency login) eRA Commons User Name			
EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)			
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY
Fudan University, Shanghai, China	B.S.	1996	Biophysics
Fudan University, Shanghai, China	M.S.	1999	Neuroscience
University of Texas-Houston/MD Anderson Cancer Center, TX	Ph. D.	2003	Human & Molecular Genetics
Harvard Medical School, MA	Post-doc	2003-2007	Genetics & Genomics

A. Positions and Honors.**Positions and Employment**

1999-2000 Graduate Research Assistant, Human Genetic Center, University of Texas-Houston
 2000-2002 Rosalie B. Hite Fellow, University of Texas -MD Anderson Cancer Center
 2002-2003 Graduate Research Assistant, Center for Genome Information, University of Cincinnati
 2003-2007 Post-doctoral associate, Department of Genetics, Harvard Medical School
 2007-Present Assistant Professor, Department of Bioengineering, University of California at San Diego

Honors

2000-2002 Rosalie B. Hite Fellowship in Cancer Research, UT- MD Anderson Cancer Center
 2003 Sowell-Huggins Scholarship in Cancer Research, UT- MD Anderson Cancer Center

B. Selected peer-reviewed publications (in chronological order).

1. Akey JM, Zhang K (joined first author), Xiong M, Doris P, Jin L. (2001) The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. **Am. J. Hum. Genet.** 68:1447-1456.
2. Akey DT, Akey JM, Zhang K, Jin L. (2002) Assaying DNA methylation based on high-throughput melting curve approaches. **Genomics** 80: 376-384.
3. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. **Am. J. Hum. Genet.** 71: 1227-1234.
4. Akey JM, Zhang G, Zhang K, Jin L and Shriver M. (2002) Interrogating a high-density SNP map for signatures of natural selection. **Genome Res.** 12: 1805-1814.
5. Chen Z, Zhang K, Zhang X, Yuan XH, Yuan ZY, Jin L and Xiong M. (2003) Comparison of gene expression between metastatic derivatives and their poorly metastatic parental cells implicates crucial tumor-environment interaction in metastasis of head and neck squamous cell carcinoma. **Clin. Exp. Metastasis** 20: 335-342.
6. Akey JM, Zhang K, Xiong M, and Jin L. (2003) The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. **Mol Biol Evol.** 20: 232-242.
7. Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L. (2003) Randomly distributed crossovers may generate block-like pattern of linkage disequilibrium: An act of genetic drift. **Hum. Genet.** 113: 51-59.
8. Zhang K, and Jin L. (2003) HaploBlockFinder: haplotype block analyses. **Bioinformatics** 19:1300-1301.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

9. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, **Zhang K**, Mitra RD, Church GM. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. **Science** 309:1728-1732.
10. **Zhang K**, Zhu J, Shendure J, Porreca GJ, Aach JD, Mitra RD, Church GM. (2006) Polony haplotyping of individual human chromosome molecules. **Nature Genetics** 38:382-387.
11. **Zhang K**, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM. (2006) Sequencing genomes from single cells by polymerase cloning. **Nature Biotechnology** 24:680-686.
12. **Zhang K**, Lott ST, Jin L, Killary AM. (2007) Fine mapping of the NRC-1 tumor suppressor locus within chromosome 3p12. *Biochem. Biophys. Res. Commun.* 360:531-538
13. Kowalchuk GA, Speksnijder AG, **Zhang K**, Goodman RM, van Veen JA. (2007) Finding the needles in the metagenome haystack. *Microb. Ecol.* 53:475-85
14. Porreca PJ, **Zhang K** (jointed first author), Li JB, Xie B, Austin D, Vassallo SL, Leproust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J. (2007) Multiplex amplification of large sets of human exons. *Nat. Methods* 4:931-6.

C. Research Support

Completed Research Support

None

Pending Research Support

None

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.
Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**

NAME Jae Keith Joung		POSITION TITLE Associate Professor of Pathology (HMS) Associate Chief of Pathology for Research (MGH)	
eRA COMMONS USER NAME eRA Commons User Name			
EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)			
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY
Harvard College	A.B.	1987	Biochemical Sciences
Harvard University	Ph.D.	1996	Genetics
Harvard Medical School	M.D.	1996	Medicine
Massachusetts Institute of Technology	(Post-doc)	2001	Biology

A. Positions and Honors**Positions and Employment:**

1996-1999 Resident in Clinical Pathology, Massachusetts General Hospital (MGH), Boston, MA
 1996-1999 Clinical Fellow in Pathology, Harvard Medical School (HMS), Boston, MA
 1997 Chief Resident in Clinical Pathology, Massachusetts General Hospital, Boston, MA
 1998-2001 Post-doctoral Fellow, Howard Hughes Medical Institute/Massachusetts Institute of Technology, Cambridge, MA
 1999-2000 Research Fellow in Pathology, Massachusetts General Hospital, Boston MA
 2000-2001 Co-Director, Molecular Diagnostics Laboratory, Massachusetts General Hospital, Boston, MA
 2000-2004 Instructor in Pathology, Harvard Medical School, Boston, MA
 2000- Assistant Pathologist, Massachusetts General Hospital, Boston, MA
 2003- Member, Center for Cancer Research, Massachusetts General Hospital Cancer Center, Boston, MA
 2004-2008 Assistant Professor of Pathology, Harvard Medical School, Boston, MA
 2004- Faculty member, Ph.D. program in Biological and Biomedical Sciences, Division of Medical Sciences, Harvard Medical School, Boston, MA
 2006 Associate Director, Molecular Pathology Unit, Massachusetts General Hospital, Boston, MA
 2007- Director, Molecular Pathology Unit, Massachusetts General Hospital, Boston, MA
 2007- Member, Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA
 2008- Associate Professor of Pathology, Harvard Medical School, Boston, MA
 2009- Associate Chief of Pathology, Pathology Service, Massachusetts General Hospital, Boston, MA

Honors:

1987 Graduated *magna cum laude*, Harvard College
 1989-1996 Medical Scientist Training Program (M.D.-Ph.D. Program) appointment, Harvard Medical School
 1989-1994 Life and Health Insurance Medical Research Fund, M.D.-Ph.D. Scholarship Award
 1990 & 1993 Distinction in Teaching, Derek Bok Center for Teaching and Learning, Harvard University
 1996 Bernard N. Fields Prize in Microbiology and Molecular Genetics (for Ph.D. thesis), Harvard Medical School
 1998-2000 Howard Hughes Medical Institute Postdoctoral Research Fellowship for Physicians
 2000 Board Certification in Clinical Pathology, American Board of Pathology
 2004-2005 Reviewer, Cancer Diagnostics and Treatments Special Emphasis Panel [CDT SEP- ONC 12] study section, National Cancer Institute, National Institutes of Health

B. Peer-reviewed publications (in chronological order):

Original Articles:

1. Ruben S, Perkins A, Purcell R, **Joung K**, Sia R, Burghoff R, Haseltine WA, Rosen CA. Structural and Functional Characterization of Human Immunodeficiency Virus *tat* Protein. *Journal of Virology* 1989; 63: 1-8.
2. **Joung JK**, Le LU, Hochschild A. Synergistic activation of transcription by the E. coli cAMP receptor protein. *Proc. Natl. Acad. Sci. USA*, 1993; 90: 3083-7.
3. **Joung JK**, Koepp DM, Hochschild A. Synergistic Activation of Transcription by Bacteriophage λ cI Protein and E. coli cAMP Receptor Protein. *Science* 1994; 265: 1863-6.
4. **Joung JK**, Chung EH, King G, Yu C, Hirsh AS, Hochschild A. Genetic strategy for analyzing specificity of dimer formation: *Escherichia coli* cyclic AMP receptor protein mutant altered in its dimerization specificity. *Genes and Development* 1995; 9: 2986-2996.
5. Dove SL, **Joung JK**, Hochschild A. Activation of prokaryotic transcription through arbitrary protein-protein contacts, *Nature* 1997; 386: 627-630.
6. **Joung JK**, Ramm EI, Pabo CO. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions, *Proc. Natl. Acad. Sci. USA* 2000; 97: 7382-7387.
7. Hurt JA, Thibodeau SA, Hirsh AS, Pabo CO, **Joung JK**. Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection, *Proc. Natl. Acad. Sci. USA* 2003; 100: 12271-12276.
8. Thibodeau SA, Fang R, **Joung JK**. An optimized high-throughput β -galactosidase assay for bacterial cell-based reporter systems, *Biotechniques* 2004; 36: 410-415.
9. Nguyen-Hackley DH, Ramm E, Taylor CM, **Joung JK**, Dervan PB, Pabo CO. Allosteric Inhibition of Zinc-Finger Binding in the Major Groove of DNA by Minor-Groove Binding Ligands, *Biochemistry* 2004; 43: 3880-3890.
10. Serebriiskii IG, Fang R, Latypova E, Hopkins R, Vinson C, **Joung JK**, Golemis EA. A combined yeast/bacterial two-hybrid system: development and evaluation, *Mol Cell Proteomics* 2005, 4: 819-826.
11. Vallet-Galy I, Donovan KE, Fang R, **Joung JK**, Dove SL. Repression of phase-variable *cupA* gene expression by H-NS-like proteins in *Pseudomonas aeruginosa*, *Proc. Natl. Acad. Sci. USA*, 2005, 102: 11082-11087.
12. Meng X, Smith RM, Giesecke AV, **Joung JK**, Wolfe SA. A counter-selectable marker for bacterial-based interaction trap systems, *Biotechniques*, 2006, 40: 179-184.
13. Giesecke AV, Fang R, **Joung JK**. Synthetic protein-protein interaction domains created by shuffling Cys2His2 zinc fingers, *Molecular Systems Biology*, 2006, doi: 10.1038/msb4100053
14. Wright DA, Thibodeau-Beganny S, Sander JD, Winfrey RJ, Hirsh AS, Eichtinger M, Fu F, Porteus MH, Dobbs D, Voytas DF, **Joung JK**. Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly, *Nature Protocols*, 2006, 1: 1637-1652.
15. Sander JD, Zaback P, **Joung JK**, Voytas DF, Dobbs D. Zinc Finger Targeter (ZiFIT): An Engineered Zinc Finger/Target Site Design Tool, *Nucleic Acids Res.*, 2007, 35: W599-W605; doi: 10.1093/nar/gkm349.
16. Meng X, Thibodeau-Beganny S, Jiang T, **Joung JK**, Wolfe SA. Profiling the DNA-binding specificities of engineered Cys2His2 zinc finger domains using a rapid cell-based method, *Nucleic Acids Res.*, 2007, 35: e81.
17. Cornu TI, Thibodeau-Beganny S, Guhl E, Alwin S, Eichtinger M, **Joung JK**, Cathomen T. DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases, *Molecular Therapy*, 2008, 16: 352-8.
18. Pruett-Miller SM, Connelly JP, Maeder ML, **Joung JK**, Porteus MH. Comparison of Zinc Finger Nucleases for Use In Gene Targeting in Mammalian Cells, *Molecular Therapy*, 2008, 16: 707-717.
19. Ramirez CL, Foley JE, Wright DA, Muller-Lerch F, Rahman SH, Cornu TI, Winfrey RJ, Sander JD, Fu F, Townsend JA, Cathomen T, Voytas DF, **Joung JK**. Unexpected failure rates for modular assembly of engineered zinc fingers, *Nature Methods*, 2008, 5: 374-375.
20. Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichtinger M, Jiang T, Foley JE, Winfrey RJ, Townsend JA, Unger-Wallace E, Sander JD, Muller-Lerch F, Fu F, Pearlberg J, Gobel C, Dassie J, Pruett-Miller SM, Porteus MH, Sgroi DC, Iafrate AJ, Dobbs D, McCray PB, Cathomen T, Voytas DF, **Joung JK**. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification, *Molecular Cell*, 2008, 31: 294-301; PMC2535758.

21. Fu F, Sander JD, Maeder ML, Thibodeau-Beganny S, **Joung JK**, Dobbs D, Miller L, Voytas DF. Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc finger arrays, *Nucleic Acids Res.*, 2009, 35: D279-D283; doi:10.1093/nar/gkn606.
22. Sander JD, Zaback P, **Joung JK**, Voytas DF, Dobbs D. An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins, *Nucleic Acids Res.*, 2008, doi:10.1093/nar/gkn962.
23. Foley JE, Yeh J-R Y, Maeder ML, Reyon D, Sander JD, Peterson RT, **Joung JK**. Rapid Mutation of Endogenous Zebrafish Genes Using Zinc Finger Nucleases Made by Oligomerized Pool ENgineering (OPEN), *PLoS ONE*, 2009, 4: e4348; doi:10.1371/journal.pone.0004348.
24. Townsend JA, Wright DA, Winfrey RJ, Fu F, Maeder ML, **Joung JK**, Voytas DF. High Frequency Modification of Plant Genes Using Engineered Zinc Finger Nucleases, *Nature*, 2009, *in press*.

Review Articles/Book Chapters:

1. Hochschild A, **Joung JK**. Synergistic activation of transcription in *Escherichia coli*. *Nucleic Acids and Molecular Biology* 1997; 11: 101-114.
2. Serebriiskii I, **Joung JK**. Yeast and Bacterial Two-hybrid Selection Systems for Studying Protein-protein Interactions. In: *Protein-Protein Interactions: A Molecular Cloning Manual*, E.A. Golemis, editor. Cold Spring Harbor Laboratory Press, 2001, pp. 93-142.
3. **Joung JK**. Identifying and modifying protein-DNA and protein-protein interactions using a bacterial two-hybrid selection system. *J. Cellular Biochem.* 2001, Supp 37: 53-57.
4. **Joung JK**, Lewandrowski KB. Laboratory Safety: An Overview. In: *Clinical Chemistry – Laboratory Management and Clinical Correlations*, KB Lewandrowski, editor. Lippincott Williams & Wilkins, 2002, pp. 40-50.
5. Thibodeau SA, **Joung JK**. An improved strategy for constructing “designer” Cys2His2 zinc finger proteins, *Discovery Medicine* 2003, 3: 32-35.
6. Hirsh AS, **Joung JK**. Engineered Cys2His2 Zinc Finger DNA-Binding Domains, *Gene Therapy & Regulation* 2004, 2: 191-206.
7. Giesecke AV, **Joung JK**. A bacterial two-hybrid system for studying and modifying protein-protein interactions. In: *Protein-Protein Interactions: A Molecular Cloning Manual*, 2nd ed., EA Golemis & P Adams, editors. Cold Spring Harbor Laboratory Press, 2005, pp. 195-216.
8. Thibodeau-Beganny S, **Joung JK**. Engineering Cys2His2 Zinc Finger Domains using a Bacterial Cell-Based Two-Hybrid Selection System, *Methods in Molecular Biology*, vol. 408: *Gene Function Analysis*, M. Ochs, editor. Humana Press, 2007, pp. 317-334.
9. Giesecke AV, **Joung JK**. The Bacterial Two-Hybrid System as a Reporter System for Analyzing Protein-Protein Interactions, *CSH Protocols*, 2007, doi:10.1101/pdb.prot4672
10. Cathomen T, **Joung JK**. Zinc-Finger Nucleases - The Next Generation Emerges, *Molecular Therapy*, 2008, 16: 1200-1207.

C. Research Support

Ongoing Research Support:

R01 GM069906 (Joung) 08/01/04-06/30/09

NIGMS/NIH

Studies of NRSF/REST zinc finger-DNA interactions

This grant award is aimed at studying protein-DNA interactions mediated by the NRSF/REST transcription factor.

Role: PI

R01 GM072621 (Joung) 04/01/05-03/31/09

NIGMS/NIH

Zinc Finger Protein-Protein Interactions

This grant award is aimed at performing structure, function, and design studies of protein-protein interactions mediated by zinc finger domains.

Role: PI

R24 GM078369 (Joung) 03/01/07-02/28/11

NIH/NIH

DNA-binding specificities of Cys2His2 zinc fingers

This grant award is aimed at developing a probabilistic recognition code for Cys2His2 zinc fingers.

Role: PI

Private Source

04/01/07-03/31/09

Private Source

This pilot award is aimed at developing non-allele-specific zinc finger nuclease reagents for gene targeting of CFTR exon 10.

Role: Co-I

R21 RR024189 (Walker) 07/01/07-06/30/09

NCRR/NIH

Zinc Finger Targeting of *C. elegans* Genes

This grant award is aimed at developing zinc finger nuclease-stimulated gene targeting for use in *C. elegans*.

Role: Co-I

Private Source

04/01/08-03/31/10

Private Source

This grant is aimed at developing zinc finger nucleases that can mediate highly efficient insertion of transgenes into a safe harbor locus in human cells.

Role: Co-I

R21 HL091808 (P.B. McCray) 04/01/08-03/31/10

NIH/NHLBI

Allele-Specific Repair of the CFTR DeltaF508 Mutation using Zinc Finger Nucleases

This grant is aimed at developing allele-specific zinc finger nucleases that recognize the delta F508 CFTR mutation and to test their use for gene correction of this mutation in human cells.

Role: Co-I

Private Source

01/01/09-12/31/09

This grant supports efforts to utilize oligonucleotide library synthesis (OLS) product for engineering combinatorial zinc finger libraries

Role: PI

European Research Council (Isalan) 03/01/09-02/28/11

Subcontract

Engineering customized zinc finger DNA-binding domains

This subcontract supports the engineering of customized zinc finger DNA-binding domains for use in system biology experiments

Role: Collaborator

Completed Research Support:

5K08DK02883 (Joung) 09/01/00-06/30/05

NIH/NIDDK

Structural and Functional Characterization of WT1

This grant award is aimed at identifying physiologic nucleic acid and protein interaction partners of WT1.

Role: PI

Program Director/Principal Investigator

Church, George M.

RESOURCES

FACILITIES: Specify the facilities to be used for the conduct of the proposed research. Indicate the project/performance sites and describe capacities, pertinent capabilities, relative proximity, and extent of availability to the project. If research involving Select Agent(s) will occur at any performance site(s), the Laboratory:

Facilities and Resources:

The Harvard/MIT/BU intellectual environment is excellent for multidisciplinary, collaborative and functional genomics research. The Church Laboratory provides some of the glue with students from all three universities and a location in three adjacent buildings at the heart of the HMS campus: 1) The New Research Building is home to the Genetics department";

2) The Seeley-Mudd Building is home to the Harvard Institute of Proteomics (HIP), the Harvard Institute of Chemistry and Cell Biology (ICCB), and the Lipper Center for Computational Genetics. 3) The Thorn Building of the BW Hospital Genomics & Bioinformatics Center. Harvard has recently made considerable endowment commitments to the above and the University-wide Center for Genomics Research. We work closely with our departmental Biopolymers facility, which has a staff of five, departmental computer facility with a staff of four. We have direct computer network and CAD links to the HMS machine shop, which coordinates with several other university and commercial machining and design facilities.

Clinical:

Animal:

Computer:

The Group has an extensive computer facilities and CAD-PAM software for design of DNA constructs. Computers are connected via LAN to the HMS campus network for access to scientific literature.

Office:

The PI's office space is in the New Research Building, at 77 Avenue Louis Pasteur.

Other:

Five -20C freezers and two -80 freezers

5 high voltage (500V to 6000V) power supplies (Biorad and EC)

MAJOR EQUIPMENT: List the most important equipment items already available for this project, noting the location and pertinent capabilities of each.

- 2 Affymetrix Chip Scanners (HP & MD) and fluidics stations
- 1 Microarrayer prototype (Anorad stages), 150 slide capacity, 16 piezoheads (GeSim)
- 1 microarray scanner (General Scanning 5000)
- 1 Automated DNA and protein sequencers, synthesizers and related items
(ABI 3700, 377, 373S, 391, 1000S, 394, 380B, 270A, 477A, 430A, 420A, 130A)
- 1 FPLC and Phast systems (Pharmacia)
- 1 LCQ HPLC-MSn Ion Trap mass spectrometer (Finnigan)
- 1 Storm Fluorimager with 29 exposure plates (Molecular Dynamics)
- Numerous PCR machines with 96, 384-well, and slide heads (MJR)
- 1 Microfluidics development platform (Caliper)
- 5 -20 C freezers and two -80 freezers
- 7 low-speed centrifuges and ultra-centrifuges (IEC, Sorvall, Beckmann)
- 2 Oscilloscopes and 2 electrophysiological amplifiers 70 femtoamp rms (Axon)
- 1 micropipette puller and microforge (Narishige)
- 5 high voltage (500V to 6000V) power supplies (Biorad and EC)
- 2 Ultra-thin gel Direct Transfer Electrophoresis (HMS shop, Cykal)
- 1 96-pin array oligonucleotide synthesizer Primer Station 960 (IAS & HMS)
- 3 electrotransfer devices (Polytech)
- 1 pulsed-field CHEF boxes (Genplex)
- 1 UV crosslinker (HMS shop)
- 1 Capillary array electrophoresis prototype (HMS shop)
- 1 Laser-induced fluorescent 4-color capillary electrophoresis (ABI 310)
- 2 DEC alpha file servers running Ultrix
- 1 dual Intel PII, RAID level 5 based Linux fileserver
- 15 computers running under WinNT, 10 Linux, 6 Linux&NT, 5 MacOS
- 1 Silicon Graphics Octane computer
- 1 Linux Celeron Cluster (Beowulf-type) for parallel & associative processing
- 1 Terabyte tape jukebox server running Arkeia
- 1 Confocal Microscope (Biorad)
- 1 Automatic film processor
- 1 Bioflo 3000 mammalian and microbial cell culture chemostat (New Brunswick)
- 1 EPICS ALTRA flow sorter with Autoclone multiwell plates option (Beckman-Coulter)
- 1 M5 plate reader (Molecular Devices)
- 1 KBiosciences high-capacity thermal cycler
- 3 Danaher Motion Dover Polonator DNA sequencers

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

RESOURCES

FACILITIES: Specify the facilities to be used for the conduct of the proposed research. Indicate the project/performance sites and describe capacities, pertinent capabilities, relative proximity, and extent of availability to the project. If research involving Select Agent(s) will occur at any performance site(s), the biocontainment resources available at each site should be described. Under "Other," identify support services such as machine shop, electronics shop, and specify the extent to which they will be available to the project. Use continuation pages if necessary.

Laboratory:

Dr. Daley's laboratory encompasses 5 laboratory modules (4 benches per module) plus chemical room, instrument rooms, tissue culture space totaling 3500 sq ft in the new Karp Family Research Building at Children's Hospital. The space is split between the 7th floor (controlled by Children's Hospital) and the 5th floor (controlled by the Brigham and Women's Hospital through a lease arrangement). In addition, he has access to shared equipment and facilities within the Division of Hematology/Oncology, the Karp Family Research Building at Children's Hospital, and the Dana Farber Cancer Institute. The laboratory is fully equipped for routine protein biochemistry, molecular biology, and tissue culture as described in this proposal.

Clinical:

N/A

Animal:

N/A

Computer:

Dr. Daley has at least 8 computers and 2 printers dedicated to the laboratory. Dr. Daley's lab members also have access to the extensive computer facilities of the Hematology Division, networked to research computing facilities both at Brigham and Women's Hospital and to the Internet. In addition to routine computer equipment and software, the Division has a digital camera, a high-resolution slide scanner, a slide maker and two high quality color printers for making slides and artwork.

Office:

Dr. Daley has 180 square feet of office space, and additional cubical space for an administrative assistant.

Other:

MAJOR EQUIPMENT: List the most important equipment items already available for this project, noting the location and pertinent capabilities of each. Within Division of Hematology/Oncology: X-Omat; phosphorimager (Molecular Dynamics); cell sorter (Becton Dickinson FACS Calibur); ultracentrifuges (one Beckman L60, two Beckman L70s, one Beckman TL-100 tabletop model); beta and gamma counter (Beckman LS3801 and 5500B); digital camera with gel scanner; HPLC (Waters) and FPLC (Pharmacia) equipment; Leica histochemistry equipment (including microtomes for standard and frozen sections and an automatic slide stainer); 2 liquid N2 freezers; lyophilizer (Virtis); sonicator (Heat Systems); UV crosslinker (Stratagene); SpeedVacs (Savant); three research photomicroscopes (Nikon Exlipse E800 equipped for brightfield, darkfield, phase, epifluorescence and differential interference contrast microscopy, with 4x/10x/20x/40x/60x/100x Plan-Fluor objectives, a high resolution (1996x1450) digitizing color scanner, image enhancement software; Zeiss Axioskop photomicroscope equipped for brightfield, phase and epifluorescence; Zeiss Axiovert 35 inverted photomicroscope.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

RESOURCES

FACILITIES: Specify the facilities to be used for the conduct of the proposed research. Indicate the project/performance sites and describe capacities, pertinent capabilities, relative proximity, and extent of availability to the project. If research involving Select Agent(s) will occur at any performance site(s), the biocontainment resources available at each site should be described. Under "Other," identify support services such as machine shop, electronics shop, and specify the extent to which they will be available to the project. Use continuation pages if necessary.

Laboratory:

Dr. Kun Zhang has a laboratory (Room 406) of ~900 sqft in the Powell-Fotch Bioengineering Hall. It is equipped with a fume hood and appropriate gas, water, and vacuum lines. The laboratory can comfortably accommodate eight full time experimental biologists.

Clinical:

Animal:

Computer:

The Department of Bioengineering has a linux cluster of 104 nodes. Each node has two CPU and 2GB of memory.

Office:

Dr. Kun Zhang has an office of ~600 sqft, which is divided by three rooms (402, 403, 404) for the PI, post-docs and graduate students.

Other:

MAJOR EQUIPMENT: List the most important equipment items already available for this project, noting the location and pertinent capabilities of each.

Housed in the laboratory: four PCR thermocyclers, one real-time thermal cycler (Biorad Chromo4), two AirClean PCR hood, one NanoDrop spectrophotometer, one Biorad GelDoc XR, one Eppendorf Vacufuge concentrator, two incubators, one -80C ultrafreezer, two -20 freezer, one 4C frigerator, one Eppendorf 5804R benchtop refrigerated centrifuge, two Eppendorf microfuges.

Housed in the shared Biotech Core in the Bioengineering department: Bio-rad HRLC AS-100 automatic sampling system, Hitachi F-2000 fluorescent spectrofluorometer, Molecular Device E-max plate reader, Molecular Dynamics Storm 840 scanner, two Innova 4000 incubator shakers, Stragagene 2400 UV stratalinker, one Scotsman ice machine, one tissue culture room.

Housed in UCSD BioGem core facility: three Illumina Genome Analyzer II (GA II, one partially owned by the Department of Bioengineering). There are four additional GA II in other individual labs in UCSD, and one ABI SOLID.

Principal Investigator/Program Director (Last, First, Middle):

Church, George M.

RESOURCES

Laboratory:

The Joung laboratory is located in the MGH Molecular Pathology Unit on the 6th floor of Building 149, a MGH research facility located in Charlestown of which Dr. Joung is the Director. The Joung lab occupies ~1000 square feet of research space and is fully equipped for molecular biology and biochemical studies including a dedicated room with space for RNA isolation and tissue culture work. The Molecular Pathology Unit occupies ~13,000 square feet of space including ~10,000 square feet of laboratory space and ~3,000 square feet of offices and conference room space. These superbly equipped research facilities are designed to maximize interactions and collaboration among the nine independent investigators that occupy this space. The Unit has extensive shared core equipment and resources (see detail below). In addition, the Unit is located one floor below the MGH Center for Cancer Research, a comprehensive molecular biology research facility. Dr. Joung is a member and principal investigator in the MGH Center for Cancer Research and therefore has privileges with all of this Center's associated core facilities.

Clinical: N/A

Animal: N/A

Computer:

The Joung laboratory is equipped with 8 PC workstations. One additional PC workstation is present in the PI's office. The Joung lab network is directly connected to the hospital network which provides high-speed access to the Internet. Two laser printers are connected to the Joung lab network. The Joung lab workstations run on Windows XP Professional.

Office:

Dr. Joung's ~100 square foot office is located adjacent to the laboratory. Dr. Joung's workstation is connected to a color laser printer. Secretarial and grants management support is provided by the Molecular Pathology Unit. Fax machines and photocopiers are also available for use in the Unit

Other:

The MGH DNA Sequencing Core and the Dana-Farber/Harvard Cancer Center DNA Sequencing Core both provide low cost, high quality automated sequencing services with a rapid turnaround time (Dr. Joung has full access to both of these facilities). The Molecular Pathology Unit partially funds and utilizes dishwashing services available through the MGH Cancer Center. Members of Dr. Joung's lab have access to multiple autoclaves in this dishwashing facility (located on the same floor as the Unit).

MAJOR EQUIPMENT: List the most important equipment items already available for this project, noting the location and pertinent capabilities of each.

Equipment in Dr. Joung's lab includes a FACScan machine (and associated Macintosh G4 workstation) for flow cytometry experiments, a complete Micronics system for labeling, storing, and tracking 2-D bar-coded individual tubes containing either DNA or bacterial glycerol stocks in compact 96-well format, two tissue culture hoods, two Jouan CO2 incubators, clinical centrifuge, ABI 392 DNA synthesizer, multiple refrigerators and freezers (-20 and -80 C), high voltage power supplies, equipment for gel electrophoresis, three PCR thermal cyclers, a digital imaging system (BioRad Gel Doc), BioRad plate reader, ELISA plate washer, dedicated shakers and incubators for bacteriological work, BioRad Gene Pulser II electroporation apparatus, thermal shaker unit, and a hybridization oven. Shared equipment in the Molecular Pathology Unit (to which Dr. Joung has full access) includes an ABI 7900 HT Sequence Detection System machine (for quantitative real-time PCR and RT-PCR), BioRad phosphorimager with dedicated computer, Beckman spectrophotometer, two fume hoods, cold room, two high speed centrifuges, two ultracentrifuges, two fume hoods, a cold room, liquid nitrogen storage systems, scintillation counter, microscopes, and a dedicated workstation linked to the Center for Cancer Research Microarray Core network that is fully equipped with software for analyzing the results of microarray experiments.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

2. Specific Aims

The goal of our proposed Center for the Causal Transcriptional Consequences of Human Genetic Variation (CTCHGV) is to develop methods that will identify and characterize cause-effect relationships between human genome sequence variation and transcriptional networks, with specific focus on cis transcription. Recent genome wide association studies (GWAS) involving many human cohorts have improved our knowledge of human genetic variation and its relationship to human physiology and disease. Yet these developments are only early steps towards the detailed causal understanding of how genetic variation relates to phenotype needed to translate this knowledge to effective clinical practice. This is particularly so for variation in non protein coding regions which comprise 99% of the genome and which obey few known rules. As of this writing, 315 GWAS studies have uncovered 1439 SNPs that associate with ~200 human traits with $p < 1e-5$ (59), 95% of which are in non coding regions. Most of these are tag SNPs in linkage disequilibrium with possibly causative SNPs as yet unidentified or even assayed in their subjects. The tag SNPs are typically common variations and the contribution to human health and disease of common vs. rarer variations is still debated. Meanwhile, ongoing sequencing of diverse populations and the growing number of sequenced individual human genomes yield an ever-increasing number of previously unseen and rare point, indel, and rearrangement variations. To move from association to cause in a manner that is not complicated by variation rarity, population sampling, and sequencing depth, CTCHGV will develop and demonstrate innovative techniques that establish cis variants' causal status by systematically and precisely varying cis sequences at single-nucleotide resolution using synthetic biology techniques, so that the effects of these variations on cis gene transcriptional level can be observed directly (Aim 1). Data generated by these methods will directly explain many GWAS findings of associations between cis variations and expression levels (40, 132, 165), and will enable refinement of hypotheses for disease causation where GWAS finds associations between cis regulatory loci and disease or phenotype. Moreover, to assist such refinement, CTCHGV will extend application of these new methods to human induced Pluripotent Stem cells (iPS) in order to make its methods able to explore the impact of cis variations in diverse human cell types representing different tissues (Aim 2). To achieve scalability in its methods for discerning sequence causality by systematically examining combinations of variations, CTCHGV will develop methods that operate with small samples of cells, including methods that assay many individual cells. Here, to determine causal cis variants requires only that the transcription of one gene—the cis gene—be assayed in small samples and in single cells. To extend beyond this and observe the systematic effects of variations, CTCHGV will therefore also develop new methods for obtaining transcriptome level information in single human cells, including both dispersed cells and in-situ structured tissues (Aim 3). Finally, CTCHGV will develop a number of innovative basic enabling technologies to achieve the scale and control over DNA synthesis and cell handling required to meet the goals above (Aim 4). As these technologies will have great impact and wide utility in biological research, CTCHGV will develop them with general usage in mind, in an open-source manner, and in collaboration with our many academic and business partners.

Aim 1: We will develop and demonstrate novel methods that identify and characterize natural cis variations that directly affect transcriptional activity in individual humans based on direct modification and testing of combinations of variants in gene regulatory regions in cell lines, and that can be applied to thousands of genes.

1.1: We will develop and demonstrate novel, high-efficiency methods to create human cell populations containing combinations of natural variations in gene regulatory regions, focusing on zinc-finger nuclease (ZFN)-mediated recombination of externally generated altered insert libraries, and direct modification of human cells using oligo-based methods.

1.2: We will demonstrate the identification of specific sets of variations that affect cis gene transcription by engineering many combinations of variations and directly observing their effects on transcription, and also by novel methods of assaying complex populations of combinatorially modified cells at a single-cell level.

1.3: We will assess the extent to which cis variants identified as causing altered transcript expression may operate through alternative mechanisms such as differential expression of RNA isoforms, differential transcript degradation, copy number variations, and epistatic marks.

1.4: We will analyze the relationship between our methods and results and those of Genome Wide Association Studies and characterize their complementary insights into the effects of variation.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Aim 2: We will adapt and extend Aim 1 methods to function in human induced Pluripotent Stem cells (iPS) and then use iPS to characterize the effect of cis regulatory region variations in a variety of derived cell types that represent different human tissues. We will engineer "marked allele" human iPS that are heterozygous in all exons of many genes that will enable analysis of allele-specific transcriptional and splicing effects in diverse cell types.

2.1: We will combine Aim 1 methods with automated techniques for iPS generation and maintenance to enable exploration of iPS with altered cis regulatory regions.

2.2: We will differentiate iPS generated in Aim 2.1 into diverse cell types that represent distinct human tissues and characterize the cell type-specific consequences of cis-regulatory variations.

2.3: We will engineer human iPS with "marked alleles" for 10-50 genes and demonstrate their use by characterizing allele-specific transcription and splicing in multiple tissues.

Aim 3: We will develop novel single-cell in-depth transcriptome assays scalable to millions of individual cells simultaneously in both structured tissues and dispersed cell samples, subject to sequencing capacity. These methods will be used to explore systematic transcriptional effects of genetic variations in different human cell types.

3.1: We will develop and optimize methods that pipeline in-situ single-cell cDNA synthesis to next generation sequencing in ways that preserve cell identity and that can be applied in parallel to 100s to 1000s of cells. We will investigate multiple techniques in support of these methods, including cell bar-coding, in-situ cell sequencing, and single-molecule in-cell sequencing, characterize their performance and limits, and select one for continued development and application.

3.2: We will use these single cell transcriptomics capabilities to characterize the transcriptional state differences in cells bearing artificial and natural variant combinations from Aim 1, and from cell types developed from iPS from different genetic backgrounds.

Aim 4: In support of Aims 1-3, we will develop innovative and widely applicable methods for high-throughput synthesis of long DNA constructs, highly efficient homologous recombination in human cells, and highly multiplexed single cell handling that enables sorting based on morphology.

4.1: We will develop a platform that integrates DNA synthesis and sequencing and uses sequence information to assure synthesis of DNA constructs with extremely low error rates.

4.2: We will improve ZFN-mediated homologous recombination in human cells by engineering a comprehensive zinc-finger archive, by developing novel methods of delivering ZFNs into cells, and by developing a "segmental genome replacement" strategy.

4.3: We will develop new high-throughput cell handling and sorting capabilities that can incorporate morphology information in addition to optical signals generated by markers, and which can operate on live cells.

Beyond the five years of our Center, we foresee the innovations we develop being applied at large scale by partner academic research centers such as the Broad Institute, as well as their adoption and further development by sequencing and synthesis companies with which we have close relationships (see Data and Materials Dissemination Plan). Our approaches will help biomedical research in general move beyond population-based associations to causal understanding. Their application to individual humans vs. populations will be critical for developing the knowledgebase required to promote and evaluate the effectiveness of personalized medicine.

Our world-class team has expertise in all the areas required for success in this project and has a track record of impact and innovation. Professor George Church (Harvard Medical School), CTCHGV's proposed director, has several times developed innovations that exhibited improvement factors of 10 or more in scale or power compared to contemporaneous commercial collaborators. Indeed, Professor Church led a prior Molecular Genomics and Imaging CEGS (MGIC) that consistently developed improved sequencing methods ~2 years ahead of commercial efforts which later adapted many of our innovations: Under him, MGIC demonstrated his initial polymerase colony (polony) methods in 2003 (116, 117), versions of which are now widely used commercially (Illumina, ABI), while in 2005 MGIC developed sequencing by ligation (155), which is now in use in ABI SOLiD. Another example is in DNA synthesis, where he has led the way in synthesis and use of complex oligo mixtures cleaved from arrays for large construct assembly and targeted sequencing, and where in the course of four years he has advanced from 4000 90-mer to 54000 150-mer oligo arrays (100,

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

174). Dr. J. Keith Joung (Harvard Medical School, Massachusetts General Hospital) is a leading expert on the development of zinc-finger nucleases for human cell engineering and gene targeting. He is the leader and co-founder of the Zinc Finger Consortium (<http://www.zincfingers.org/>), which was established to ensure and to promote continued research and development of engineered zinc finger technology. The Consortium is committed to developing a zinc finger engineering platform that is robust, user-friendly, and freely available to the academic scientific community. Dr. George Q. Daley's (Harvard Medical School, Children's Hospital, HHMI) work has transformed the field of stem cell development and differentiation. Recipient of numerous awards, including the first NIH Director's Pioneer Award, as well as major awards from the American Philosophical Society, Society for Pediatric Research, Burroughs Wellcome Fund, and the Leukemia and Lymphoma Society of America, Dr. Daley's work focuses on functional hematopoietic and germ cell elements from ES cells, and the genetic mechanisms that predispose to malignancy. Dr. Daley's lab was one of the first three world-wide to derive human iPS cells, and the first to produce a repository of patient-specific iPS cells (from 10 different disease conditions). Professor Kun Zhang (UCSD) developed innovative methods for long range haplotyping, single cell genome sequencing, targeted sequencing, and measurement of allele-specific expression, as a post-doc in the Church Lab, where he was a member of the MGIC team. He is currently working with Professor Church on methods for targeted exon sequencing in connection with an NHLBI grant (HLB08-004). In addition to these key personnel, CTCHGV will have additional support from experts in GWAS, genome wide screens in human cells, and companies offering sequencing support (see Letters of Support from David Altshuler, Steven McCarroll, Robert Plenge, Steven Elledge, and Complete Genomics, among others).

3. Background and Significance

3.1 Determining the causal consequences of natural human genetic variations: Rapid developments over the past 8 years led quickly from the completion of draft human genomes (97, 180) to the identification of common human genetic variations (146) and haplotypes (68), and then to the use of these variations to identify loci associated with human traits and disorders through Genome Wide Association Studies (GWAS) (188). GWAS, conceptualized early in the Human Genome Project (144), has become the leading method for linking genetic variation with human traits: As of this writing, 315 GWAS studies have uncovered 1439 SNPs that associate with ~200 human traits with $p < 1e-5$ (59), 95% of which are in non coding regions. By their nature, GWAS do not identify the causal mechanisms linking variations with traits but only the associated variations themselves, but these may provide important clues about the causal networks underlying the traits. However, limitations of GWAS are frequently noted (4): The identified variations may not themselves be in the causal network for a trait but may only be tag SNPs in linkage disequilibrium with causative variations, and in practice GWAS can only be used to detect common SNPs with moderate effect sizes that can be cost-effectively genotyped in some few thousands of individuals. As a result, efforts are underway to make rarer mutations analyzable by GWAS (4, 67, 126), and to expand the GWAS paradigm to identify variations specifically associated with gene expression level (notably *via* expression Quantitative Trait Loci (eQTL) (23, 32, 40, 46, 165) and allele-specific expression (ASE) (152)) and to investigate the effects of Copy Number Variations (111, 164). These advances will refine our understanding of the causative networks underlying traits, but in the end they may still fall short of identifying the specific variations that are causal. This is because GWAS' power to dissect co-occurring human variations and genetic and environmental backgrounds is ultimately constrained by the spectra of natural human variation, population structure, and linkage disequilibrium that have been shaped by human evolutionary history. CTCHGV proposes to develop and demonstrate technologies that will enable such genetic covariates to be dissected in the specific domain of variations that affect cis gene expression levels. These technologies will be based on direct engineering of combinatorial modifications to cis regulatory region genotypes, followed by direct observation of the effects of these modifications on cis gene transcription levels. While these methods will not dissect direct links between variations and disease, they will help refine hypotheses concerning disease causation where GWAS finds associations between disease and regulatory variations. To assist this analysis, CTCHGV will extend its technologies to human induced Pluripotent Stem cells (iPS) so that the causal impacts of regulatory variation can be examined in diverse human cell types representing different tissues. In this way, CTCHGV will provide an essential complement to GWAS: While GWAS take us from phenotype to associated locus by deeply leveraging natural human variation, CTCHGV methods will dissect the roles of variations that are below its limen.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

The methods that will be developed by CTCHGV in its five years of operation will specifically focus on the characterizing the effects of cis variations on transcriptional level (Aim 1) and some of their downstream consequences (Aim 3). Eventually, follow-on work will require methods that analyze the broad range of other molecular mechanisms, including effects of variations on RNA splicing, protein expression level and translational control, as well as the effects of epigenetic variation and imprinting. Here Aim 1.3 will assess a selection of these effects in specific contexts in an effort to quantify their importance and chart a path forward.

3.2 Single cell transcriptomics: Acquiring detailed information about the transcriptional state of individual cells is critical for understanding the development of complex organisms and the functions of structured, differentiated tissues. The leading methods available today involve the isolation of individual cells by disaggregation/dilution or microdissection, extraction of cell contents by lysing or micropipetting, followed by synthesis and subsequent amplification of cDNA to levels at which it can be assayed via RT-PCR or microarray (44, 83, 94). Though these methods have been applied successfully in numerous studies, they are subject to important limitations: (i) Cell isolation procedures have limited scalability and, where tissue cells must be disaggregated and diluted, may destroy structural information about a cell's location in a tissue. (ii) The need for very high pg-to- μ g amplification of single cell mRNA increases the potential for introducing biases into sample transcript abundances. To enable transcriptomes to be obtained for large numbers of individual cells will require addressing both of these limitations. In common with most technology development aimed at increasing throughput, (i) is best addressed by miniaturization and parallelization of operations on single cells. Microarrays require large amounts of starting material and so exacerbate problem (ii). Researchers are increasingly turning to next-generation sequencing to supplant microarrays generally, resulting in such methods as RNA-Seq and PMAGE (84, 185), and this has recently been applied to single cells (169). While this will partly alleviate (ii), application of these methods in parallel to many single cells as required by (i) will still require further advances in sequencing technology.

The ideal solution would be one in which cell transcriptomes could be sequenced *in-situ*; this would confine and parallelize sequencing within single cell compartments, and also allow structured tissues to remain intact so that cell locations and relationships are preserved. The challenge to sequencing technology is that, at present, most next-generation sequencing methods must create and operate with spatially distinct, compactly localized amplicons of sample sequences dispersed on planar surfaces (154, 156); thus, new methods would be needed to create and interrogate these amplicons in existing cell volumes. (Note that the localized amplification needed by sequencers is distinct from the high gain in-solution amplification of cDNA required by microarrays; for multiplex single-cell sequencing, the latter should be avoided as it not only increases sequencing costs but may also re-introduce amplification biases.) However, possibilities exist (a) for creation of localized amplicons in stacks of very thin sections of cells that can be sequenced *in-situ*, and (b) for transcripts of individual cells to be labeled *in-situ* in a way that preserves cell identity so that these transcripts could be dispersed, locally amplified, and sequenced *ex-situ* as usual and re-assigned to their cell of origin. Also (c) single molecule sequencing methods (45, 55) can in principle avoid the need for local amplification, but interrogating individual transcripts in existing cell volumes is still beyond the reach of these methods as they are configured today, especially for (45). Aim 3 will explore these possibilities. Aim 3 work on (a) will greatly scale up prior methods developed in our own and other labs on *in-situ* localized amplification and detection of RNAs (163, 202) in the direction of detecting and quantifying many thousands of RNA species in cells. Assuming cells of $\sim 10\mu\text{m}$, a 1 mm^2 section of tissue that is one cell thick would contain $1\text{e}4$ cells. If each cell is assumed to contain $\sim 2\text{e}5$ coding transcripts, sequencing 50bp tags of all coding transcripts in these cells would require $2\text{e}9$ reads and $1\text{e}11$ bp. Illumina has recently announced that its Genome Analyzer will be able to produce $\sim 1\text{e}11$ bp of sequence per run by the end of 2009 (66). As recent trends of increasing sequencing throughput per dollar are expected to continue, CTCHGV does not see sequencing capacity and cost as inherent limitations to the ability to sequence transcriptomes of all cells in a tissue sample of this size by the end of the five years of our proposed center. However, as the availability of sequencing capacity on this order is likely to be a practical consideration for many research projects for some time, CTCHGV will develop options for assaying of targeted transcriptome subsets vs. complete transcriptomes in single cells that will give researchers the ability to allocate available sequencing capacity as best fits their needs.

3.3 induced Pluripotent Stem Cells (iPS): Full understanding of the effects of natural variations in humans requires exploration of their impacts in different tissues. Expression Quantitative Trait Loci (eQTL) found by observing quantitative differences in gene expression in multiple individuals are numerous and highly heritable (40, 119, 150, 182), but these examine only limited numbers of tissue types, and it requires a large

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

number of tissue samples to reach adequate statistical power due to their generally weak individual effect in addition to measurement noise and other confounding factors (32, 78). Although eQTL mapped from different tissue types overlap (23, 24, 46, 121, 149), many regulatory pathways are known to be tissue- and cell type-specific. To address these limitations, the Genotype-Tissue Expression Project (125) has been launched to collect various tissue types from a large number of subjects. However, collection of diverse human tissues using surgical and tumor specimens is complex and affected by sampling and processing artifacts (32), and these issues must be overcome for a very large number samples to detect the weak effects of most eQTL.

By contrast, iPS technology (133, 167, 168, 195) allows biomedical researchers to derive cells representative of numerous tissues and cell-types *in vitro* from a single common source—a superficial skin biopsy—making iPS cells a powerful platform for studying individual differences in gene regulation without limitation of tissue or cell type. Although it is unclear whether conclusions based on iPS-derived tissue cell types can be generalized to primary tissues, their use offers undeniable practical advantages in terms of ease of cell type access, controllable purity of cell type vs. the complex composition of primary tissues, and ability to work directly with human cells vs. other species. Admittedly, this strategy may also face challenges, including epigenetic alterations in gene expression among iPS clones, heterogeneous cell differentiation, and changes in genome-wide DNA methylation (108, 113, 114, 161); moreover, there may be random mono-allelic gene expression in iPS clones, as reported for EBV-transformed lymphoblasts (50, 137). Finally, typical iPS reprogramming is accompanied by numerous random viral integration events (~20 per clone), possibly affecting the expression of nearby genes; however, we anticipate eliminating this variable by generating and studying transgene-free iPS cells (cf. (194)). We will explore use of allele-specific expression (ASE) to assess *cis* regulation in iPS-derived cells. This strategy helps control the effect of experimental variations on gene expression, which function predominantly *in trans* (102, 159), by using one of the expressed alleles as an internal control. Using our highly quantitative ASE measures (Preliminary Results, 4.2), we find that ASE is largely stable over the sources of variation described above. iPS differentiation can thus be expected to yield additional cell type-specific ASE that is not captured by use of adult somatic cell lines alone. On that basis we propose to use ASE analysis to observe and map *cis*-acting regulatory variation using human iPS cells (Aim 2).

3.4 Synthetic technologies in human cells: Decades of research have given us the tools to make multiple precise changes to the genomes of cells of model organisms, to the point where the synthesis and assembly of standardized parts for engineering complex pathways has become the defining goal of the new field of *synthetic biology* (43, 79, 174). While synthetic biology is spawning many useful applications in lower organisms, a key goal is to make its technologies operate efficiently in human cells where they can be used to analytically dissect the mechanisms underlying human traits and disease, and to implement pathway repairs that modulate disease and deliver them into targeted human cells via gene therapy. Efficient application of these synthetic technologies to human stem cells and iPS is a particularly important goal because such cells are self-renewing and therefore have potential for permanent therapeutic effect, and because they can be used to generate many cell types (cf. (54)). Motivated by these objectives, researchers have made considerable progress in engineering human cells to the point where gene therapy clinical trials have been performed or are under consideration for over 20 human disease areas (2, 3). The many key challenges that remain can be divided into a set that relate to improving the ability to apply synthetic technologies to human cells generally, and to a set that relate to achieving clinical success. While the latter involves critical problems such as efficient access and targeting of only specific cell types within the human organism, and identification of the genetic targets that must be altered to modulate disease, here we focus on the former—specifically, on the technical challenges of engineering human cells accurately and efficiently without making unwanted changes, assuming their accessibility. Three related areas can be identified that will receive focus and development by our proposed Center: (i) Because of the large size and complexity of the human genome, it is necessary either to create and introduce large fragments of DNA into human cells, or many smaller fragments of DNA into the cells, to achieve a desired result. This entails that large DNA fragments or many small DNA fragments must be generated with minimal error either by direct synthesis or by extracting and modifying the relevant regions of the native cells' genome. (ii) Efficient delivery systems are needed to introduce these fragments of DNA into the target cells. Currently, modified viral vectors are the most efficient delivery systems, but they can only typically package 5-25kbp (31) of DNA, and retroviral systems can randomly integrate DNA into the genome. While random integration is particularly concerning clinically for gene therapy, and is quite possibly the cause of development of cancers in the otherwise most successful gene therapy trial (19), it remains a concern in a general engineering context simply because it creates unwanted modifications in the genome. (iii) Where a

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

delivery simply introduces DNA into the cell and relies on native mechanisms to integrate it, such as electroporation, nucleofection (<http://www.amaxa.com/>) or lipofection, the efficiency of desirable low-error replacement of targeted DNA by homologous recombination (HR) and/or gene conversion (GC) is low compared to error-prone non-homologous end joining (NHEJ) and random integration. Estimates of native HR:NHEJ efficiencies vary from 1:30 to 1:40000 (191).

The state of the art in these three areas can be summarized. (i) While large DNA constructs can be synthesized *de novo*, in part due to developments by the Church Lab (174) whereby DNA oligonucleotides (oligos) synthesized on arrays can be assembled into units of 1000s of nucleotides, most human genome engineering will entail making single or multiple small changes precisely in native human DNA. Many labs have developed methods for making small changes in human DNA directly in human cells using variously-designed oligos and small DNA fragments (64, 166), while the Church Lab is nearing completion on a project that uses oligo-based methods to replace all 314 instances of the TAG stop codon in the *E. coli* genome with TAA stop codons (see Preliminary Results 4.4). Both oligo-based methods are relevant for CTCHGV and will be pursued in Aim 1 (section 5.1.1). Meanwhile, *de novo* synthesis of large DNA constructs will be relevant to generation of zinc finger nucleases (see (iii) below and section 5.4.1). (ii) Transfer of BAC-size fragments of DNA to human cells by modified bacteria has been reported by several labs and presents advantages of supporting transfer of very large DNA constructs with high integrity compared to viral and chemical/electrical methods (52, 98, 124, 135). (iii) The Joung Lab within CTCHGV and others have developed methods for design and use of zinc-finger nucleases (ZFNs) for targeting specific genomic locations for highly efficient HR. Here, fusion proteins are constructed between sets of three or four zinc-finger DNA binding domains that are designed to recognize particular nucleotide sequences, plus a type-IIS endonuclease domain (usually *FokI*) so that, when introduced into a cell, the fusion proteins dimerize and introduce a double stranded DNA break (DSB) at a selected unique genome location. The DSB is then repaired with high efficiency by HR vs. NHEJ using homologous DNA introduced into the cell. Targeted gene replacement efficiencies as high as 29% have been reported using these methods. A consortium of academic laboratories (The Zinc Finger Consortium; <http://www.zincfingers.org>) led by co-investigator Keith Joung has developed "open-source" reagents, protocols, and software that enable researchers to engineer their own ZFNs (48, 107, 134, 175). A company (Sangamo Biosciences, <http://www.sangamo.com/index.php>) has also developed their own platform for engineering custom zinc-finger proteins and access to this technology is available through Sigma-Aldrich at a price per zinc finger nuclease pair of \$25,000 (115, 134).

Within Aims 1 and 4, CTCHGV plans to both use these methods and make extensive improvements in them in support of proposed Center goals. The prospect of success is high not only because CTCHGV labs have already made key contributions to these developments (see above), but because CTCHGV provides a focus for specifically improving the engineering aspects of human cell synthetic technologies apart from the clinical aspects of gene therapy: In particular, CTCHGV's concentration will be on modifying potentially large (~100kb) human DNA constructs and introducing them into human cells and using ZFNs to efficiently drive HR, and also on using oligos to make multiple targeted small modifications to human DNA, in a research context that does not require simultaneously addressing tissue accessibility, cell targeting, or clinical impact. However, CTCHGV success will be immediately translatable to clinical gene therapy, both via CTCHGV's improved engineering methods, generally, and because in Aim 2 CTCHGV also commits to making these methods work in human iPS (see 3.4 above; also cf. (54)).

3.5 Personalized medicine: The premise of personalized medicine is that genomic information can identify individuals with different profiles of disease risk, response to treatments, or susceptibility to side effects, and thus be used to stratify individuals to optimal treatment and surveillance regimes (13). Genetic risk screening, pharmacogenetics, and expression profile assays are already in use, with 1422 clinical genetic tests for disease (177) and 23 drugs with pharmacogenetic testing information on their labeling available as of this writing ((158), Table 5). While translation from research to clinic of such genomic tests depends on many factors such as education of practitioners and patients and the development of low cost and reliable assays, the success of personalized medicine will ultimately depend on their efficacy in predicting disease and improving treatment. The CTCHGV Aims have potential to improve such efficacy in two ways: First, by refining understanding of the causal consequences of human variation, CTCHGV methods will help identify with greater precision than GWAS which variations carried by any particular individual have consequences for disease and treatment, enabling improved accuracy and sensitivity of genomic tests. CTCHGV methods also have potential to themselves be the basis of new kinds of personalized tests that could directly inform

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

decisions about disease monitoring or treatment. For instance, we can envision creating iPS from hair follicles or skin fibroblasts from an individual with a family history of cancer, developing disease-relevant tissue cells from the iPS, and then creating populations of cells bearing single gene deletions covering hundreds to thousands of genes mimicking loss of heterozygosity, to see if phenotypes or transcription profiles related to the cancer appear; and we could subsequently help prioritize available treatments by identifying those that best relieve these phenotypes or profiles.

4. Preliminary Results

As CTCHGV research will apply and significantly extend methods and expertise developed by our former Molecular and Genomic Imaging CEGS (MGIC), we begin our discussion of preliminary results by mentioning MGIC areas of achievement that will be relevant to CTCHGV goals. MGIC made important contributions to development of targeted DNA (10, 100, 139) and RNA (101) sequencing, single cell genomics (200), long range haplotyping (176, 201), image analysis (7, 8, 83), instrumentation design for automation (172), stem cells (including induced Pluripotent Stem cells (iPS); see 4.2 below), and RNA splicing (203), RNA editing (101), microRNA (181), and methylation analysis (10). CTCHGV will also leverage MGIC's considerable achievements in next generation sequencing, which not only included the development and release of the open-source, commercially available Polonator instrument (172), but in numerous methods and reagents that have been incorporated into other commercial instruments (see Data and Materials Dissemination Plan for companies with which the Church Lab within MGIC established close collaborative relationships). While selected aspects of these MGIC-related developments will be described below, we will focus mostly on preliminary results regarding other developments relevant to CTCHGV; however, see References for a list of the 45 published or pending articles produced by MGIC. We also note that MGIC's track record in training and education is relevant to CTCHGV (see Training and Minority Action Plans).

4.1 Targeted sequencing of DNA and RNA: The Church and Zhang Labs have been leaders in development of targeted sequencing based on Molecular Inversion Probes (MIPs) or padlock probes. In these methods, padlock probe oligos that can be synthesized on an array are designed with ends that hybridize specifically to regions flanking thousands of target sequences of interest. In a single reaction including the template DNA, the oligos, and both polymerase and ligase, the polymerase extends the oligo 3' ends across the target sequences until the 5' end of the oligo is reached, at which point the polymerase-extended oligo is circularized. The circles are purified and common sequences built into the oligos are used to amplify the targets and as sequencing primers. The method can be applied equally well to genomic DNA or cDNA; an illustration of the application of the method to cDNA is shown below in Figure 4.2-1.

The goal of targeted sequencing is to reduce sequencing requirements by restricting sequence feature creation and coverage to only the targeted subset of the initial sequences provided, which could be as large as an entire genome or transcriptome. Since their initial development during the MGIC Center (139), the Church and Zhang labs have carried on development of padlock probe methods as part of NHLBI grant HL08-004 with Jon Seidman of Harvard Med School, the goal of which is to develop targeted sequencing of human exomes of Personal Genome Project (136) (see 4.9 below) and Framingham Heart Study subjects. In addition to padlock probe-based methods, the HL08-004 project is also developing hybridization-based enrichment of exonic targets from fragmented genomic DNA. Ongoing optimization of padlock probes has improved their efficiency by 10,000-fold (100). Other than the NHLBI exomes, the Church and Zhang labs have developed and demonstrated targeted capture and sequencing via padlock probes of many thousands of targets in parallel for assays of several biologically important phenomena, including measurement of allele-specific expression of transcripts (see 4.2 below), measurement of variation rates in CpG dinucleotides (100), genome-wide assessment of CpG methylation levels (10), and detection of RNAs subject to RNA editing (101).

While padlock probe-based sequencing is a strong interest of the Church and Zhang labs, CTCHGV will not hesitate to employ or combine their use with other methods where this is advantageous. For instance, one advantage of padlock probes is that they can help equalize coverage of targets of different abundance. In this way, measurements of allele-specific expression (ASE, see 4.2 below) of low abundance RNA targets can be made more precise than those based directly on non-targeted sequencing of transcripts. However, this same feature of padlock probes may be a source of bias when the object is to compare the relative expression levels of different transcripts. Therefore, CTCHGV may use RNA-Seq or P-MAGE (84), or gene expression microarrays in conjunction with padlock probes, where both relative transcript abundance and ASE must be

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

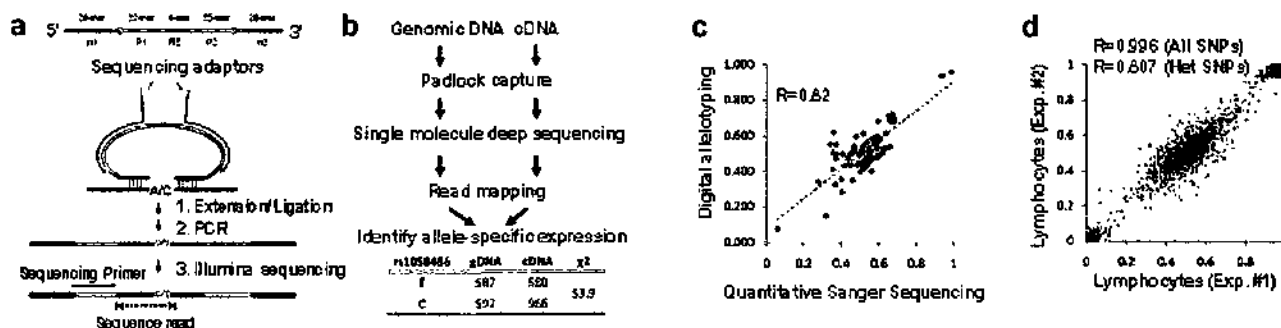


Figure 4.2-1. RNA allelotyping. (a) A schematic diagram for MIP capture and single-molecule sequencing. (b) Detection of allele-specific gene expression. (c) Comparison of allelic ratios measured by RNA allelotyping and quantitative Sanger sequencing. (d) Comparison of allelic ratios between technical replicates.

assessed (see, e.g., Specific Aim 1.3).

4.2 Allele specific expression (ASE) and long-range haplotyping: ASE: We recently adapted molecular inversion probes (MIPs) for digital quantification of RNA allelic ratios. To do this, we designed a library of 27,000 probes, each targeting a common SNP in a transcribed region (Figure 4.2-1a). All 27,000 SNPs were captured from both genomic DNA and cDNA in single-tube reactions, and the allelic ratios were determined by ultra-deep sequencing (Figure 4.2-1b). The capturing and sequencing protocols have been extensively optimized, such that the allelic ratios could be measured accurately (Figure 4.2-1c) and consistently (Figure 4.2-1d).

Haplotyping: (i) *Amplification of single human chromosome molecules.* We have extended the polymerase cloning method to amplification and sequencing of single human chromosomes. To do this, we trapped lymphocytes at the metaphase and extracted intact human chromosome molecules. The chromosome solution was then diluted to ~0.5 chromosome/reaction and amplified. The amplicons were then labeled and hybridized with a regular chromosome spread. The specific FISH signals indicated that large single chromosome fragments or intact chromosomes could be specifically amplified (Figure 4.2-2a,b).

(ii) *Improvement of Multiple Displacement Amplification (MDA) for single molecule amplification.* Our original polymerase cloning method relies on MDA, which has an inherent limitation in the representation bias. Recently, it was reported that MDA using longer randomized primers (N9) in the presence of trehalose provides more even genome coverage (131). We have confirmed the reduced bias of this method by performing MDA on single human lymphocytes followed by genotyping with Illumina Infinium chips. The N9 primer exhibited slower amplification kinetics than the conventional N6 primer, probably due to lower priming efficiency. We found that a new LN9 primer containing partial locked nucleic acids has higher amplification efficiency and a lower level of background amplification (Figure 4.2-2c,d). We are currently assessing LN9-based genome coverage using Illumina genotyping and next-gen shotgun sequencing.

(iii) *Post-normalization protocol.* Amplification bias on single template DNA molecules is unavoidable and leads to requirements for increased sequencing. We recently found that biased sequencing libraries can be normalized

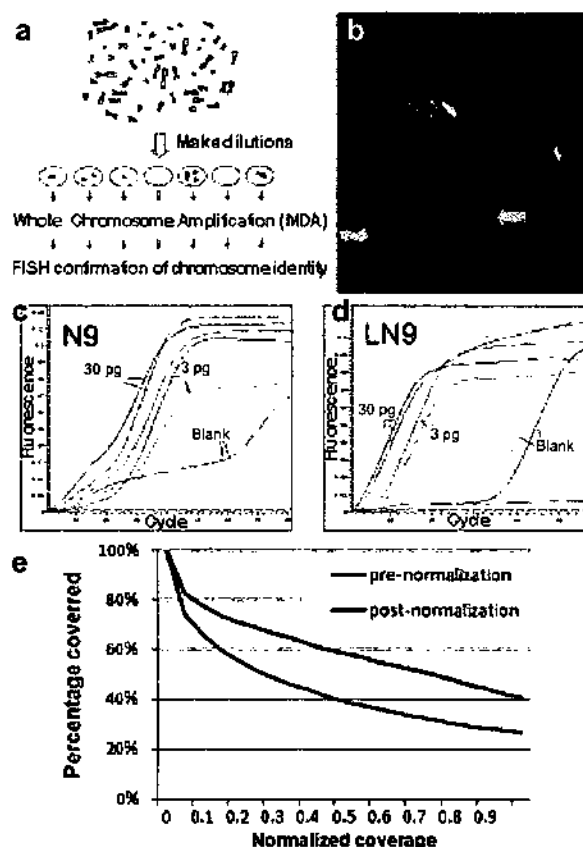


Figure 4.2-2. Polymerase cloning of human chromosome molecules. (a) MDA is performed on limited dilutions of human metaphase chromosome molecules. (b) FISH hybridization confirmed that one amplicon (purple) was from chromosome 6 and another amplicon (green) was from chromosome 19. (c, d) Real-time amplification curves with the N9 and LN9 primers. (e) Reduction of representation bias with post-amplification normalization.

The N9 primer exhibited slower amplification kinetics than the conventional N6 primer, probably due to lower priming efficiency. We found that a new LN9 primer containing partial locked nucleic acids has higher amplification efficiency and a lower level of background amplification (Figure 4.2-2c,d). We are currently assessing LN9-based genome coverage using Illumina genotyping and next-gen shotgun sequencing.

(iii) Post-normalization protocol. Amplification bias on single template DNA molecules is unavoidable and leads to requirements for increased sequencing. We recently found that biased sequencing libraries can be normalized

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

prior to sequencing. The normalization procedure involves denaturing and slowly annealing the libraries, digestion of annealed sequences with a double-strand specific DNA nuclease, and enrichment of the single-stranded species with PCR. The percentage of sequences that have at least half of the average sequencing coverage was increased from 41% to 61% (Figure 4.2-2e).

4.3 Stem cells: The Daley Lab within CTCHGV is a world leader in stem cell research in all aspects of stem cell and iPS generation, differentiation, and analysis described in this proposal. Here we focus on Church Lab experience that is relevant to analysis of allele specific expression (ASE) in iPS and derived cells and the automation of iPS generation and maintenance procedures. In order to explore whether iPS cells and their derivatives can be used for cis-regulatory mapping, we derived iPS cells and performed allele-specific expression (ASE) analysis on pluripotent and differentiated cells. We captured 27,000 expressed exonic SNPs on 10,345 human genes using padlock probes (10, 100, 101, 139). Differential allele expression is then measured as a ratio between the numbers of the reads mapped to the two alternative alleles (reference vs. alternative allele). We found ASE patterns in iPS biological replicates to be highly reproducible. When each of these replicates was treated with trans-retinoic acid for 12 hours to induce differentiation, we observed significant changes in ASE, which we likewise observed in differentiating embryoid bodies (EBs). Despite large changes in epigenetics during iPS reprogramming (10), we find that ASE differences from primary fibroblasts are relatively small. Despite these variations, the overall ASE signature (up to 50%) was invariant among different cell types, culture conditions and cell batches (Figure 4.3-1). We estimate that 5-15% of genes may show differentiation-specific changes in ASE. Our results show conclusively that random allelic-bias and epigenetic influences are relatively small for iPS and iPS-derived cell lines, which can thus be used for reliable mapping of individual-specific cis-regulatory variants.

Because iPS-derived cell differentiation most closely mimicks embryonic development, the iPS transcriptome may not reflect the relevant expression signatures in adult tissues due to aging, tissue damage, and other factors. To trigger a diverse set of trans-acting regulators capable of teasing out cis-acting variants in adult processes, we partially reprogrammed primary fibroblasts using adenoviral pluripotency factors (pAdeno-OCT4, pAdeno-KLF4, qAdeno-MYC, qAdeno-SOX2). We found that many developmental trans-acting regulators were upregulated, particularly those for mesodermal (e.g., lymphocyte and muscle/skeletal) development. We also found that innate inflammation induced by adenoviral infection caused transcriptional changes consistent with immune system activation. Using this system, we were able to detect cis-variants that are relevant to adult medical disorders such as HIV-1 Rev binding protein, INFR2 and SWAP-70, as well as developmental ASE information also obtained from iPS cells. Our results revealed that using adenoviral reprogramming may be informative for common adult medical disorders associated with tissue inflammation.

In our first steps to optimize and automate the reprogramming process, we have begun using retroviral mono-vectors to deliver the reprogramming factors and have begun adapting iPS cell culture on microcarriers in collaboration with Global Cell Solutions, Inc. Preliminary data suggests that iPS cells and hES can be maintained independently of feeder layers while retaining pluripotency for at least 2-3 passages. We are currently attempting nucleofection (Amaxa) and retroviral reprogramming on magnetic microcarrier beads in mini-bioreactors. Preliminary results using primary fibroblasts indicate that this method may be superior to normal electroporation using trypsinized and/or suspended cells. We have been adapting the bioreactor overflow for iPS reprogramming in a high-throughput manner, which will facilitate genome-wide engineering of pluripotent cell lines used in Aim 2.

4.4 Genome engineering, synthesis of large DNA fragments and combinatorial libraries: The genome engineering and synthesis needs of CTCHGV can be achieved through three main strategies: 1. *de novo* DNA synthesis, 2. Multiplex Automated Genome Engineering (MAGE) in *E. coli* or 3. Direct Multiplex Automated Genome Engineering in Human cell lines. Importantly, each strategy is rooted in technology that was recently developed or in development in the Church laboratory, uniquely positioning us to generate

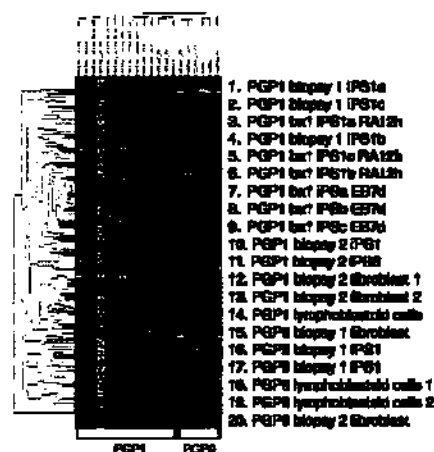


Figure 4.3-1. Hierarchical clustering of statistically significant allele-specific expression (ASE) in reprogrammed cells, showing that ~50% of overall ASE signature was invariant among different cell types, culture conditions and cell batches.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

combinatorial libraries of upstream cis variants.

4.4.1. de novo DNA synthesis: In prior work, we developed an inexpensive and high-throughput technology for large-scale DNA synthesis (174). In these experiments, we synthesized all 21 genes that

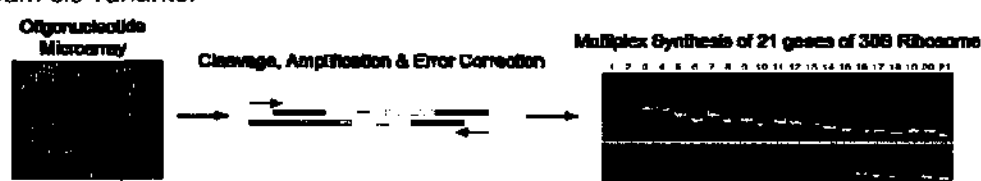


Figure 4.4.1-1 Multiplex DNA Synthesis of large DNA fragments (174)

encode the proteins of the *E. coli* 30S ribosomal subunit and mutated their DNA sequence to optimize their translation efficiency (Figure 4.4.1-1). In related work, we employed a circular assembly amplification method that significantly reduces DNA error to construct genes encoding a thermostable DNA polymerase (11). We maintain ongoing efforts to improve the fidelity and scale of DNA synthesis which will be further developed in CTCHGV Aim 4.1 and used to generate zinc finger nucleases (ZFNs) for Aim 1. Using a single 244,000 feature programmable DNA microchip, CTCHGV could generate all the ZFNs required to analyze 1000 genes in one subject (1000 genes x 5 loci/gene x 2 alleles/locus x 20 oligos/ZFN (see Overviews, sections 5 and 5.4)).

4.4.2. Multiplex Automated Genome Engineering (MAGE).

The Church Lab has pioneered the development of MAGE for large-scale programming of cells. MAGE simultaneously targets many locations on the chromosome for modification in a single cell or across a population of *E. coli* cells, thus producing combinatorial genomic diversity (Figure 4.4.2-1). In ongoing work in the Church Lab under the auspices of the Church Lab's Department of Energy Genomes-to-Life Center, we have been replacing all 314 instances of the TAG stop codon in the *E. coli* genome by TAA stop codons, thereby freeing up TAG for other possible uses. In this project, we divided the *E. coli* genome up into 32 ~145kbp segments and used MAGE to replace all coding TAGs with TAAs in each segment. At this time all segments have been completed and we are now in the process of joining them to generate a complete functioning *E. coli* genome that lacks the TAG codon. Genome construction is proceeding via a hierarchical series of conjugations of segment-bearing strains supplemented by suitable markers and selections that ensure that entire and not just partial segments are recombined. These techniques have direct relevance to methods we propose to apply in Aims 1.1 and 4.2 (see sections 5.1.1 and 5.4.2).

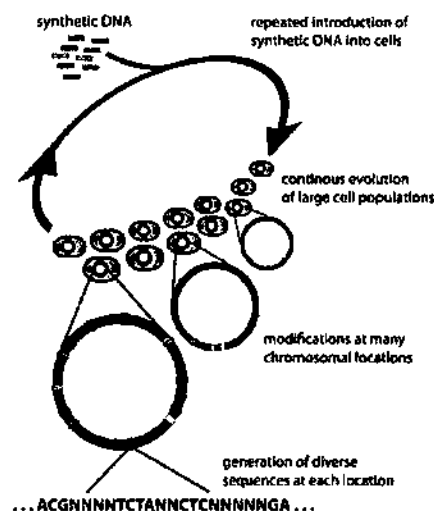


Figure 4.4.2-1 Multiplex Automated Genome Engineering (MAGE)

In another recent application, we used MAGE for large-scale programming and evolution of cells *in vivo* (184) to optimize the 1-deoxy-D-xylulose-5-phosphate (DXP) biosynthesis pathway in *E. coli* to overproduce the industrially important isoprenoid lycopene. As many as 24 genetic components in the DXP pathway were modified simultaneously using a complex pool of synthetic DNA, creating over 4.3 billion combinatorial genomic variants per day. We isolated variants with more than five-fold increase in lycopene production in less than 3 days, a significant improvement over existing metabolic engineering techniques. Since the process is cyclical and scalable, we constructed prototype devices that automate the MAGE technology to facilitate rapid and continuous generation of a diverse set of genetic changes (mismatches, insertions, deletions).

4.5 Zinc-finger nucleases (ZFNs) for improved homologous recombination (HR): OPEN (Oligomerized Pool ENgineering) is a rapid, publicly available zinc finger engineering method that was developed by the Joung lab (107) which has led academic efforts to advance engineered zinc finger technology ((134), also <http://www.zincfingers.org>). Like other combinatorial selection-based methods, OPEN identifies combinations of fingers that effectively deal with context-dependent DNA-binding effects. However, OPEN is simpler than other methods because it uses an archive of pre-selected zinc finger pools constructed to bind a variety of different 3 bp target "subsites". With the current set of zinc finger pools targeted to 66 subsites, OPEN can be used to target a sequence once every ~200 bp of random sequence. In addition, a large number of OPEN selections can be performed very rapidly – at present, two technicians in the Joung lab

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

can perform 48 selections in less than two months (M. Maeder, J. Foley, and J.K. Joung, unpublished data). OPEN is the only publicly available method which has been successfully used to create ZFNs that modify endogenous genes in human cells: Specifically, using OPEN ZFNs, the Joung lab and collaborators have used OPEN ZFN pairs to modify target sites in four endogenous human genes (*VEGF-A*, *HoxB13*, *CFTR*, and *PIG-A*) ((107) and unpublished research). Gene targeting/HR induced by OPEN ZFNs was so efficient that as many as four copies of *VEGF-A* could be modified in a single cell. In addition, the Joung lab (working with the Peterson lab at Massachusetts General Hospital) also successfully used OPEN in recent unpublished work to generate ZFN pairs for additional target sites in various endogenous zebrafish (*Tfr2*, *dopamine transporter*, *telomerase*, *HIF*, and *gridlock*) and plant (*SuRB*) genes (48).

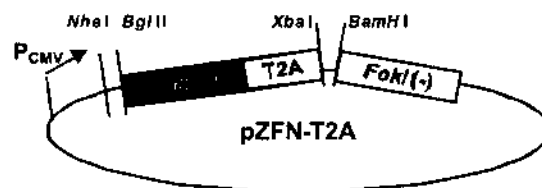


Figure 4.5-1. pZFN-T2A – dual ZFN expression vector

Direct comparisons in human cell-based assays show that OPEN is more effective and yields higher quality ZFNs than previously described “modular assembly” approaches (see Figure 2 in (107)). The higher success rate of OPEN is likely attributable to its greater sensitivity to context-dependent effects on DNA-binding among neighboring zinc-fingers that are largely ignored by modular assembly. In addition, three-finger ZFNs made by OPEN exhibit minimal toxicities in human cells compared to fully optimized four-finger ZFNs made using the complete algorithm-driven Sangamo platform (115). These findings are consistent with another recent report which demonstrated that a pair of three-finger ZFNs (made using a strategy similar to OPEN) was also no more toxic than fully optimized four-finger Sangamo ZFNs (140). Dimers of our three-finger OPEN ZFNs and Sangamo’s four-finger ZFNs should recognize 18 and 24 base pair target sequences, respectively, and, assuming full specificity, these ZFNs should be capable of recognizing genome-unique sites.

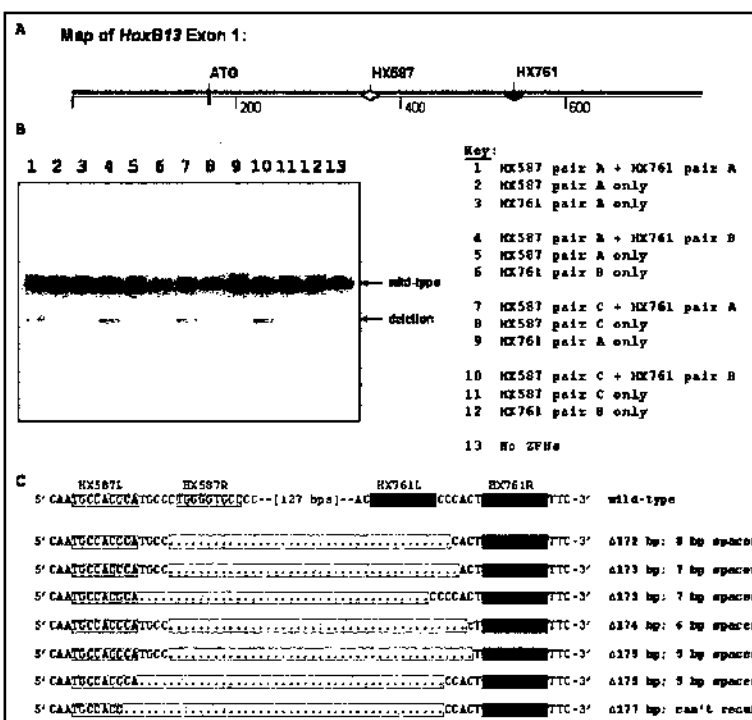


Figure 4.5-2. Dual cleavage of an endogenous *HoxB13* allele leads to deletion of intervening sequence. **A.** Map of human *HoxB13* exon 1 with sites targeted by ZFNs. **B.** Limited-cycle PCR assay of genomic DNA from cells treated with ZFN combinations of ZFNs that cut at the HX587 and HX761 sites (107). **C.** DNA sequences of deletion alleles cloned from genomic DNA of cells treated with two ZFN pairs that cleave at the HX587 (blue) and HX761 (pink) sites. ZFN half-sites are highlighted for each full ZFN site: left (L), right (R). Deletions indicated in grey.

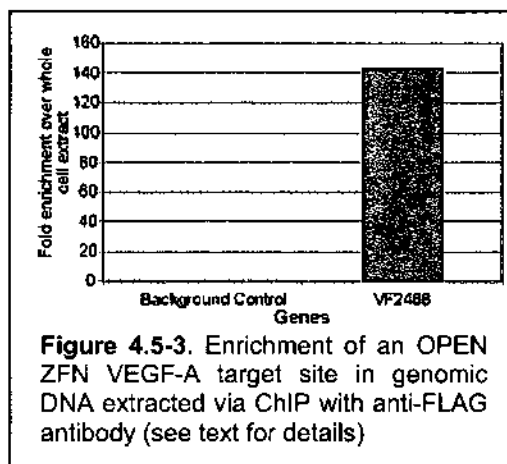
In CTCHGV Aim 4 (see section 5.4.2) we propose to improve the replacement of large segments of DNA required in Aim 1 (section 5.1.1) by using pairs of ZFNs that cut double stranded breaks (DSBs) at the flanks of the targeted segment, a strategy that should improve genome engineering capabilities generally. The Joung lab has designed and constructed a mammalian expression vector that efficiently co-expresses two ZFN monomers from a single coding transcript. In this vector, a strong CMV promoter drives expression of a single open reading frame encoding two ZFNs joined by a self-cleaving picornavirus T2A peptide (Figure 4.5-1). Previous work has shown that expression of two ZFNs joined in this way leads to efficient stoichiometric expression of the two ZFN monomers that are cleaved apart during translation and by the intervening T2A peptide (42). In our version, the two *FokI* cleavage domain coding sequences also harbor obligate heterodimer mutations which have been shown to reduce the toxicity of ZFNs due to their significant reduction of unwanted homodimer formation (115). To reduce recombination between the two *FokI* sequences, we re-coded one *FokI* monomer to make it as dissimilar to the other as

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

possible at the nucleotide level. Zinc finger arrays selected using OPEN can be excised and cloned in-frame into pZFN-T2A using the sites indicated in Figure 4.5-1. We tested the pZFN-T2A vector in human cells by using it to express a pair of ZFNs targeted to a site in the human *VEGF-A* locus (VF2468). Using a limited-cycle PCR/CEL-I enzyme mismatch detection assay that we and others have previously used to assess mutations introduced by ZFNs, we found that our vector could be used to efficiently express two ZFNs from a single vector (C. Ramirez & J.K. Joung, unpublished data).

Based on our success in simultaneously introducing two ZFN monomers into a cell, we reasoned that our pZFN-T2A vector should also enable introduction of two *pairs* of ZFNs (i.e., *four* ZFN monomers) into human cells. To demonstrate this, we used ZFN pairs targeted to two different sites in the endogenous human *HoxB13* gene—HX587 and HX761, for which the Joung lab had previously engineered pairs of ZFNs using OPEN selection (107) that each induced highly efficient non-homologous end joining (NHEJ)-mediated mutations at their respective target sites in human HEK293 cells. HX587 and HX761 are both present in human *HoxB13* exon 1 and are separated by ~180 bps (Figure 4.5-2A). We tested the hypothesis that the two pairs of ZFNs could both cleave a single *HoxB13* allele with the result that the intervening sequence might be deleted *via* rejoining of the two ends by NHEJ. After simultaneously transfecting cells with two pZFN-T2A vectors encoding pairs of HX587 and HX761 ZFNs and harvesting genomic DNA three days post-transfection, we performed limited-cycle PCR with primers flanking the two sites and found a PCR product from the doubly-transfected cells that was ~180 bp smaller than that from the wild-type allele (Figure 4.6-2B). Quantification suggests that the deletion occurs at a frequency of ~1 to 2%, although this may be an overestimate because smaller size deletion product might amplify more efficiently than the larger wild-type product. The deletion product was not visible in control experiments performed with cells transfected with only one of the two ZFN pairs or with no ZFNs (Figure 4.6-2B). Cloning and sequencing eight instances of the smaller size product revealed that they all harbored deletions of *HoxB13* sequence between the centers of the HX587 and HX761 ZFN sites, and many also exhibited additional variable length deletions on either side of each cleavage site, consistent with the hypothesis that rejoining of the two DSBs might occur via NHEJ (Figure 4.5-2C). Notably, we observed that all but one of the eight sequenced alleles should be capable of being re-cleaved with a ZFN pair comprising a LEFT HX587 and a RIGHT HX761 ZFN. Based on these results, we conclude that two OPEN ZFN pairs can efficiently cleave the same allele in human cells. In Aim 4 we will pursue the strategy of providing template DNA along with pairs of OPEN ZFNs to drive HR vs. NHEJ.

The Joung Lab (working with Bradley Bernstein at the MGH and the Broad Institute) has begun work on a method for unbiased genome-wide determination of off-target alterations to genomic DNA caused by use of ZFNs. We have proposed a very similar method in Aim 1.2 (see section 5.1.2(iv)). In this method, Chip-Seq is used to identify all binding sites in the genome of a catalytically *inactive* version of a ZFN. Subsequently, the corresponding catalytically *active* ZFN is used and targeted sequencing (Preliminary Results 4.1, above) is used to look for actual DNA alterations at these sites (for details, see 5.1.2(iv)). At this time, the Joung and Bernstein Labs have taken their procedures to the point of verifying highly specific binding of catalytically *inactive* FLAG-tagged OPEN ZFNs targeted to site VF2468 in the human *VEGF-A* promoter. The inactive ZFNs included a previously described mutated version of *FokI* (14). The inactive ZFNs were expressed as obligate heterodimers using pZFN-T2A (see Figure 4.5-1). Human K562 cells (3×10^7) were nucleofected (Amaxa) with this vector and genomic DNA harvested 24-hours post-transfection for ChIP with FLAG antibody (M2, Sigma). Initial qPCR results indicated 36-fold enrichment of the VF2468 site in the DNA from the ChIP compared with whole cell extract. ChIP DNA was then prepared for Illumina sequencing and qPCR repeated on the library DNA demonstrated 143-fold enrichment of the *VEGF-A* ZFN target site in ChIP DNA versus whole cell extract (Figure 4.5-3). The Joung Lab is awaiting results from actual Illumina sequencing.



4.6 Polonator instrument: The Polonator instrument (Figure 4.6-1) was developed as a low-cost, open source sequencing platform in our MGIC CEGS, but will not be developed as such in CTCHGV. For routine high-throughput sequencing, CTCHGV will employ available commercial platforms such as the Illumina Genome Analyzer or the Roche 454 sequencer, or look to our collaborators (see Letter of Support from

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Complete Genomics, Inc). However, CTCHGV Aims 1, 3, and 4 involve design and optimization of instrumentation for parallel single cell assays and for integrated DNA sequencing and synthesis. These can be usefully developed on the Polonator, which serves as a very general foundation for integrating flow cell-based cell handling, automated reagent handling, and integrated microscopy and image analysis. The Polonator may be used for sequencing when this needs to be integrated with new instrumentation. Here, we report that we have been developing four-color Sequencing by Synthesis (SBS) reversible terminator strategies to increase Polonator read lengths, as well as the capability to attach an ordered pattern of Rolling Circle Amplified sequences on the Polonator to increase sequencing feature density and throughput. A cyclic ligation strategy for increasing read length to 48 bases (24 from each of 5' and 3' ends) is also under development. We expect most of these improvements to be in place by the time CTCHGV funding becomes available. Our work on the Polonator, as well as on MAGE (see 4.4.2), has given us considerable expertise in instrumentation development, generally.

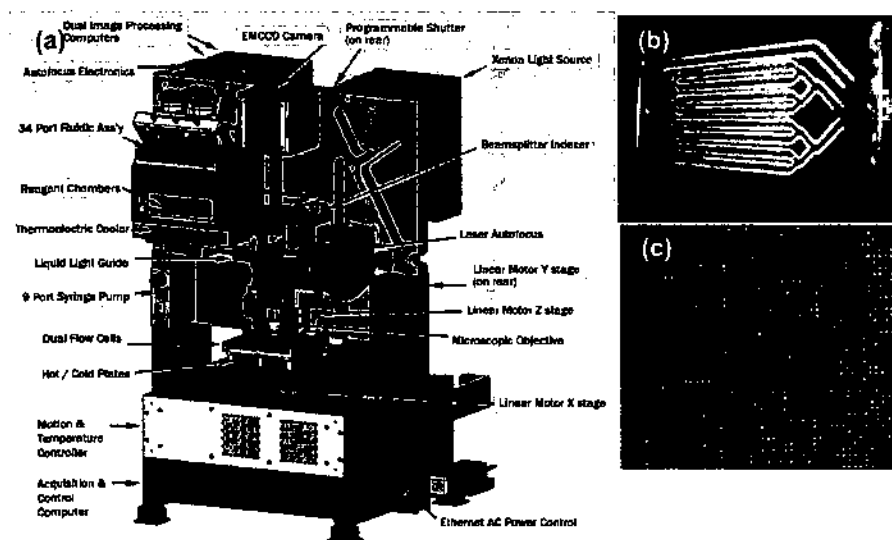


Figure 4.6-1. (a) Polonator instrument. (b) Flow cell designed and incorporated into Polonator to increase throughput and reduce runs costs by reducing reagent volumes and expenses. Overall dimensions: 150 X 70 X 8 (mm); lanes' "active area": 70 X 3.3 X 0.1 (mm) (.025 mm in testing). When loaded, the flow cell contains 0.5-1e9 1µm beads. (c) Polonies created by Rolling Circle Amplification (RCA) instead of on 1µm beads, deposited on a grid with 600 nm spot diameter and center to center spacing of 1700 nm. The image was obtained after a single Sequencing by Synthesis cycle on the Polonator using fluorescent reversible terminators (112, 190). Different colors represent the different bases incorporated during the cycle. Note that CTCHGV proposes to develop single cell transcriptomics using RCA polonies in Aim 3.

4.7 Single cell sequencing: The sequencing of a genome of an individual prokaryotic cell was accomplished by Kun Zhang while a member of the Church Lab (200), and comparable methods are used for long range haplotyping by the Zhang Lab (see section 4.2). Professor Zhang also has an R01 from NHGRI (R01HG004876) to develop a lab-on-chip device for single cell sequencing. These successes demonstrate CTCHGV capabilities applicable to single cell transcriptomics and other single cell assays (Aims 1.2, 3, 4.3).

4.8 Splice variant and methylation analysis: CTCHGV Aim 1.3 will validate the causality of cis variants identified as causing differential allelic transcription in part by exploring alternative explanations, including such phenomena as differential splicing and methylation. Church Lab experience in performing such analyses includes the following: In (203), comprehensive identification and quantification of alternative splicing for selected transcripts was performed in a single molecule gel polony framework, while (10) analyzes genome-wide methylation levels via two methods, sequencing of genomes cut by methylation-sensitive restriction enzymes, and targeted bisulfite sequencing of methylation sites (see section 4.1). Methylation is accurately measured by both measures. Versions of these procedures will be used to assess selected sets of potentially causative cis variants identified in Aim 1. RNA-Seq may be used to assess RNA splicing more globally via exon junction fragments; our development of the PMAGE (84) RNA sequencing procedure puts this within easy reach.

4.9 Personal Genome Project (PGP): The purpose of the PGP is to promote and organize the development of a set of human genome sequences and cell lines supplemented by phenotype information that can be used as research resources by the scientific community, as well as to increase public awareness of and participation in the shaping of personal genomics (28, 136). As comprehensive genomic and phenotypic

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

information is inherently identifying, a central aspect of the PGP has been the development of informed consent protocols and resources by which volunteers can educate themselves about the risks of making their data available and, at their option, consent to this (106). Towards this end, the PGP has worked closely with the Harvard Institutional Review Board (IRB) since 2004. Starting from initial approval for one participant to make available his data, the IRB has recently (February 2009) approved informed consent protocols that will enable up to 100,000 participants to volunteer their data. In the meantime, with the help of 10 initial participants (56, 60), preliminary exome and SNP data were released with phenotype information and cell lines made available through Coriell. Research community interest in the PGP has been high, with the consequence that many additional resources have been donated to the PGP, including computer equipment and software for managing the large data sets and automated processing that will be generated by the Project, as well as genomic services. Of particular interest, Complete Genomics, Inc. (<http://www.completegenomics.com/>) has committed to sequencing and making available up to 10 PGP diploid genomes (see Letters of Support).

4.10 Image and computational analysis: CTCHGV research will require sophisticated computational analysis in several key areas, including (i) image analysis, (ii) next-generation resequencing and sequence variant identification, (iii) RNA expression and ASE measurement, (iv) systems management and support for maintenance and computational analysis of large sequence and expression data sets, (v) instrumentation support, and (vi) algorithm development. The CTCHGV has substantial experience in all of these areas. (i) The Church Lab has developed sophisticated algorithms for feature calculation, morphological analysis, and classification of individual cells and cell samples (7, 8, 83). Image analysis tools have also been developed in support of gel (201, 203) and bead colonies (29, 155). (ii) The Church and Zhang labs are not only experienced with standard tools such as MAQ (99) for mapping sequence reads to target sequences and calling variants (10, 136), but have developed their own algorithms for mapping and variant calling (100), as well as for calling RNA editing sites (101). (iii) See section 4.2 for examples of CTCHGV work on ASE and RNA expression analysis. (iv) Sophisticated infrastructure and data management tools have been developed in support of the PGP (197); these tools will be available for CTCHGV. (v) Considerable software development for controlling instrumentation was built into the Polonator and will be directly usable where the Polonator is used as a framework for new CTCHGV instrumentation (see section 4.6). (vi) Initial frameworks for the algorithms that will identify causative cis variants and models of differential allelic expression are described in the Research Design Overview (section 5) and in Aim 1.2 (section 5.1.2).

5. Research Design and Methods

The goal of the CTCHGV is to develop new methods to identify cause-effect relationships between natural human genetic variations and the transcriptional states of cells, with focus on cis gene transcription. CTCHGV will do so by using synthetic biology methods to directly modify natural variations systematically in human cells to identify those combinations that result in changes in transcriptional state. We aim to develop scalable techniques that will allow combinations of variants to be analyzed for thousands of genes, and to use human induced Pluripotent Stem Cells (iPS) to generate a diverse set of cell-types. We will also develop techniques for assaying transcriptomes in many individual cells. Achieving these aims will require significant improvements to synthetic methods for generating DNA constructs for modifying human cell populations, to genetic engineering techniques for introducing and integrating these constructs into human cells efficiently, to analytic methods for simultaneously determining genetic and transcriptional state in individual human cells, and to the cell handling techniques that will integrate and automate these assays across thousands to millions of individual cells. Because CTCHGV aims only to develop and demonstrate our new technologies within our area of focus, but not to comprehensively apply them to large human populations, we plan to work with samples from a limited set of human individuals. We will use pre-existing, publicly available tissues and cell lines with potential to be transformed into iPS from HapMap, the PGP, the Framingham Heart Study, or other sources, with preference for samples for which comprehensive genome sequence is available. It will be important to start with *clonal* populations of cells from whatever source we use for reasons noted in Aim 1.2 (section 5.1.2(i)), and also to obtain diploid genome sequences of a large number of genes and their regulatory regions. Here, we will have support from our collaborator Complete Genomics, Inc (see Letter of Support).

Overview of CTCHGV Research Strategy, Numerical Targets, and Scope: The problem of identifying which natural cis variations causally affect cis gene transcription levels is important in two ways. First, knowing which specific variants causally affect transcription is important in itself (see Background and

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Significance 3.1), as this information will be relevant to understanding specific gene functions in relation to specific phenotypes. Second, the methods needed to address the problem must effectively meet several general challenges in human biological research, so that these methods will immediately have important application elsewhere. Among these issues are: (i) To identify cis causal relationships in human cells requires accurate and efficient ways of engineering them. (ii) The human organism has hundreds of identifiably different tissues, and within them, many thousands of cell types. To properly characterize cis variation requires being able to assess the spectrum of human tissue types and the complex populations of individual cells within them. (iii) The immediate transcriptional effects of causally relevant cis variations only require consideration of a single gene – the cis gene – but these may have complex and cascading downstream transcriptional effects.

The biological implications of cis variation thus require being able to track these effects in these complex populations of individual cells. (iv) To achieve these goals requires advances in technology that will enable accurate, *high-throughput* handling, manipulation, and observation of small cell populations and single cells. Broadly speaking, CTCHGV's four Specific Aims follow out these four imperatives, with Aim 1 simultaneously carrying out (i) and yielding the direct payoff of identifying specific causative cis variations for thousands of genes, while Aims 2-4 address (ii)-(iv), respectively. This overview of CTCHGV's research strategy thus starts with a careful look at Aim 1 and how it connects with Aims 2-4.

Aim 1's strategy for identifying variations that causally affect transcription levels is depicted in Figure 5-1. We start by identifying genes in subject cell lines that exhibit allele-specific expression (ASE) as measured by differential expression of indicator alleles *x* and *y* in gene coding regions (sub-Aim 1.2, section 5.1.2(i)). We then identify cis variations in putative regulatory regions of the

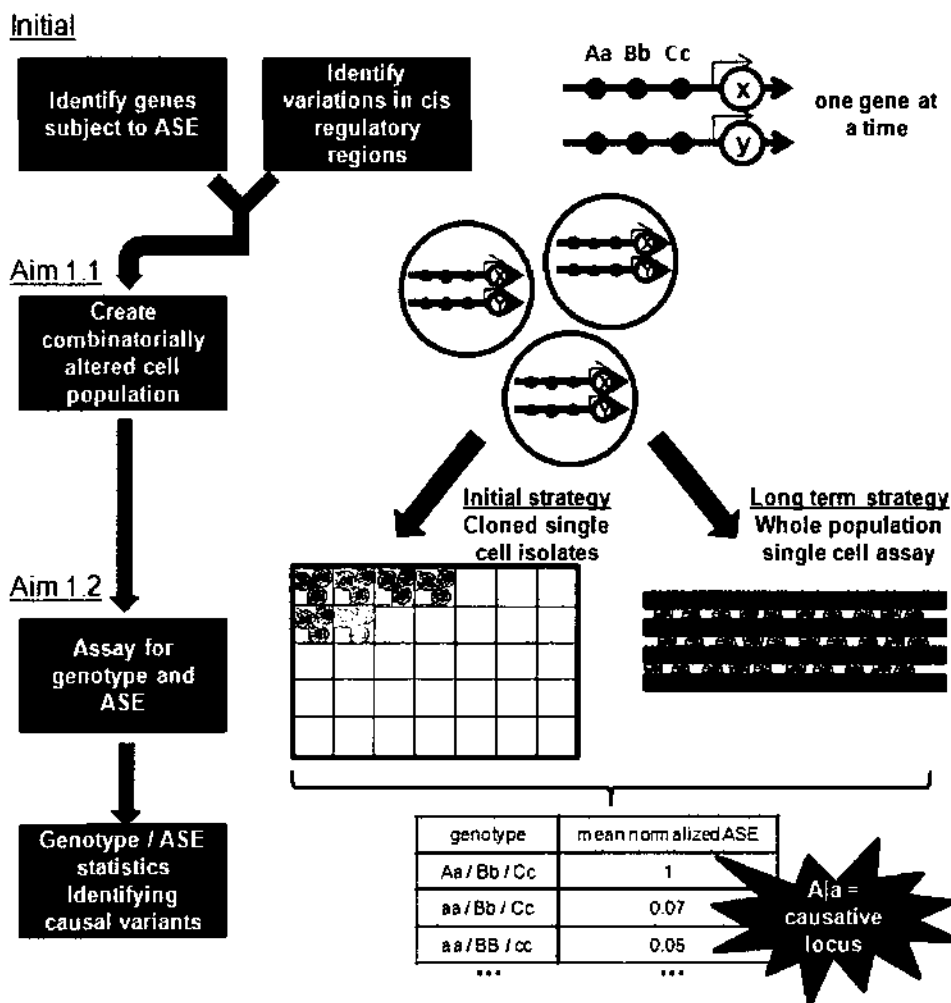


Figure 5-1: Overview of Aim 1 strategy for identifying causative cis variations. Initial Aim 1 work identifies genes subject to allele-specific expression (ASE), and, via next-gen sequence data, also identifies variations in regulatory regions, e.g., here the 100kpbs upstream region. Via Aim 1.1, cell populations are created for each gene bearing combinations of the variations identified for that gene. Via Aim 1.2, cells from this population are genotyped and assessed for ASE so that the specific loci and loci interactions that control ASE can be identified. The initial Aim 1.2 strategy will examine clonal outgrowths of individual altered cells from the population, while a longer term strategy will assay the entire mixed altered population at a single cell level. This strategy is executed one gene at a time for 100s to 1000s of genes.

genes, considering not only SNPs but also small indels and other sequence variations. Assuming ~1 variation / 1000 bp, a regulatory region such as a 100kb upstream region may contain ~100 variants. Our starting assumption is that some of these variations may actually be causes of ASE, while many of the rest are associated with it by dint of being in the same haplotype block. Our task is to identify variants that are actually

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

causal. As a first cut, we will use information on conservation and transcription factor binding sites to reduce the number of variants we will analyze to ~5 or less (sub-Aim 1.2, section 5.1.2(i)). We will then use engineering techniques developed in sub-Aim 1.1 (section 5.1.1) to, in effect, break down the haplotype block so that the transcriptional effects of these variations can be observed independently, revealing which are causal. Specifically, we will generate subject cell lines in which these five variant loci are modified to at least individually assume all haplotypic states. For instance, if one of the loci is **A|a**, with **A** cis to the **x** coding indicator allele, and **a** cis to the **y** indicator allele, we will generate cells which are **AA**, **aa**, and **Aa** in both haplotypes, and we will similarly alter the other four variant loci. To engineer these changes, we will (1) extract gene regulatory region genomic sequences from the subject cell lines, (2) alter them in *E. coli* using MAGE techniques (see Preliminary Results, 4.4), and (3) re-introduce the altered regions back into the subject lines and induce homologous recombination to replace the native regulatory region alleles using Zinc Finger Nucleases (ZFNs) that are targeted to the regions (18, 107, 115). However, we will also develop oligo-based methods which will greatly simplify and expedite altered cell line generation (sub-Aim 1.1, section 5.1.1). To generate and optimize the many ZFNs that will be needed to analyze many genes, which may include allele-specific ZFNs, we will develop high-throughput synthesis and ZFN targeting optimization methods in Aim 4 (section 5.4).

Generating cell lines with the four haplotypes per locus for each of five loci individually will entail generating 20 altered cell lines. These altered lines can be made with ZFNs one locus at a time. However, with the MAGE- and oligo-based methods we will develop (sub-Aim 1.1, section 5.1.1), we can easily generate combinatorially altered cell populations that contain all $4^5=1024$ haplotypes at once. Both strategies have advantages. If we work one locus at a time, we can identify individual loci that affect differential allele expression by themselves, and then investigate interactions systematically by manipulating specific pairs or triplets (etc.) of these loci. On the other hand, fewer ZFNs are needed to create combinatorially altered populations, and by presenting all possible combinations of the alleles at once, they also put us in position to acquire maximal information on possible interactions between the cis loci immediately—if they can be assayed efficiently (see below). A third strategy will be to create combinatorial populations and isolate sets of individually altered cells for clonal outgrowth to get a random sample of isogenic altered cell populations. We will develop each of these strategies and apply them as they best fit circumstances.

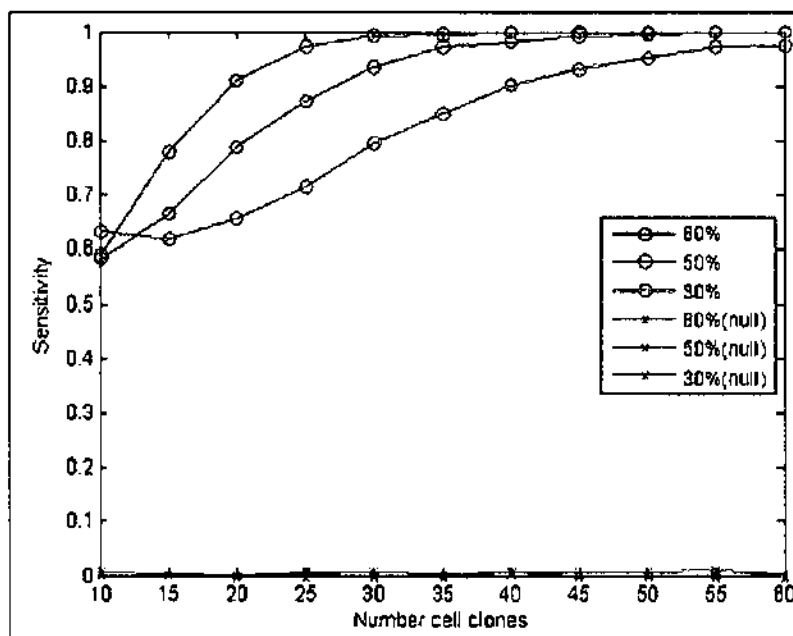


Figure 5.2: Sensitivity of detection of causality using assays of clonal altered cell lines, based on simulations involving 5 loci in a *dis* gene regulatory region. For locus **A|a**, allele **A** was assumed to cause its *cis* gene allele to be expressed at 2x the level of the **a** allele; other loci had no effect. Combinatorial cell populations were simulated in which all haplotypes occur with equal frequency whenever homologous recombination (HR) took place: HR efficiencies of 80%, 50%, and 30% were considered: Individual cells are isolated (number, X-axis), grown clonally, and assayed for genotype and gene expression level. Expression levels of the *cis* gene alleles phased with **A** and **a** were simulated subject to 20% Gaussian error. The **A|a** locus was identified as causal if gene expression level correlates with number of **A** alleles with $p < 0.01$ (corrected for 5 loci). "Sensitivity" = fraction of 1000 random simulations in which **A|a** was detected as causal. "null" lines: fraction of simulations in which non-causal locus **B|b** was identified as causal.

Having created the *cis*-locus altered cell lines, the next step is to assay them to see how specific alterations affect *cis* gene or allele expression. The most basic tests involve growing clonal populations from individual cells from the variously altered cell lines and identifying situations in which increased expression follows a particular allele of a particular locus (e.g., where gene expression is highest in **AA** lines, intermediate in **Aa** lines, and lowest in **aa** lines), or where ASE is abolished in cell lines homozygous for the locus (**AA** and

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

aa) but not in those in which the locus is heterozygous (Aa). Simulations (see Figure 5-2) show that with as few as 60 clonal populations grown from randomly selected individual cells from a combinatorially altered population, we should be able to identify which one of five cis loci cause two-fold differential allelic expression with confidence $p < .001$, with a sensitivity $\geq 97.5\%$ and false positive rate $< 1\%$, even if homologous recombination efficiency is assumed to be 30%, a value that has been achieved with current ZFN technology (18, 107, 115). We will pursue this strategy in Aim 1.2 (section 5.1.2), but we will also seek to go beyond it by looking for interactions between the cis loci. Here many models can be considered. e.g., cis loci can be additive, or upregulation might depend on specific allele phasing. These models can be distinguished by looking at the average profiles of differential expression over the various combinations of genotypes (see Table 5-1).

We can acquire information on these profiles by using clonal isolates from combinatorially altered populations. However, these operations may become cumbersome when dealing not only with larger numbers of cells, but also with the large number of genes we wish to examine. As a longer term strategy, we will therefore develop an assay that enables the combinatorially altered population for a gene to be examined as a population, wherein genotypes and ASE levels will be assayed together in millions of altered individual cells, obviating the need for isolating and growing up many single cells (details in section 5.1.2 (ii)).

The assays we perform will identify many cis variants that determine gene expression levels. These

identifications will automatically have gone deeper than associations because their participation in the cause-effect chain has been directly interrogated in cell lines with constant genetic background save for precise engineering of a few variations. However, further analysis will be needed to determine the nature of the causation. A cis variant might cause a change of expression level by altering a transcription factor binding site (170), or it might alter the methylation or histone modification profile of the regulatory region. Or, the cis variant might not actually regulate expression level but, rather, alter the splice isoform profile of the cis gene (95), so that the changes in abundance of the coding region indicator alleles x and y by which ASE is measured might be due to differential splicing of the exon containing them. In Aim 1.3 (section 5.1.3) we will assess the prevalence of a variety of such phenomena in a representative set of original and altered CTCHGV cell lines.

As noted in Specific Aims and Background and Significance (section 3.1), our proposed strategy has a close relationship with GWAS. Our methods will go beyond GWAS by identifying variants that actually cause vs. associate with phenotypes, and will also avoid GWAS constraints on effect size and allele frequency. However, our methods will themselves be limited to finding variants that specifically cause changes in expression level, compared to GWAS which finds associations with disease and phenotype. However, because GWAS frequently identify associations in non-coding regions, variations found by our methods to be causative of differential expression will generate and refine hypotheses stemming from GWAS associations. In this proposal, we make these relationships with GWAS explicit in two ways. First we use GWAS to help prioritize genes and variants that will be analyzed by our methods. Second, in Aim 1.4 we close the loop by assessing what it would take for GWAS to discover the effects we find without our methods. This analysis must, by its nature, focus on GWAS that examine expression levels vs. phenotypes, but these comparisons will illuminate how GWAS and the techniques we develop will complement each other.

The analysis of cis variants provided by Aim 1 will be constrained in several dimensions. While some scope limitations are described below, others will be addressed in other Aims. Because of the large amount of engineering that will be performed on cell lines in Aim 1, we will employ tractable cell lines that tolerate the conditions of engineering, and this constraint will limit the universe of expression profiles (including ASE profiles) under study. However, in Aim 2 (section 5.2), we will generate induced Pluripotent Stem Cells (iPS)

#A / #B	haplotypes	M1	M2	M3	M4	Table 5-1: Profiles of ASE by genotypes for four models of cis variant interaction involving two cis variant loci (A/a vs. B/b), assuming cells from a completely random combinatorial population of all haplotypes. Models M1 AB necessary and sufficient to upregulate the cis allele. M2 A alone necessary and sufficient to upregulate cis allele. M3 A and B additively and equally upregulate their cis alleles. M4 A upregulates cis allele relative to a, while B modifies A such that having A and B in cis causes an additional 50% increase in upregulation of A vs. a. ASE measurement assumes that AA and aa yield zero ASE values and Aa has a positive value regardless of haplotype. Mean ASE levels are given normalized relative to maximum possible ASE level for any genotype. #A / #B: Genotypes given by the numbers of A and B alleles respectively. Genotypes homozygous for both loci are uninformative for ASE and not shown. Haplotypes consistent with each shown genotype are given.
2/1	AB Ab	1	0	.5	0	
1/2	AB aB	1	1	.5	1	
1/1	AB ab Ab aB	.5	1	1	.83	
1/0	Ab ab	0	1	.5	.67	
0/1	aB ab	0	0	.5	0	

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

with alterations developed in Aim 1, which will enable us to explore their impact in iPS-derived cell types representing diverse human tissues. The alterations generated in Aim 1 will also be developed and analyzed one gene at a time (but are scalable to many genes). In Aim 2, we will apply Aim 1 techniques to develop complexly altered iPS in which *many genes* are altered at once. In its focus on cis variant causation, Aim 1 will mainly look at expression levels of one gene in relation to the variants it manipulates—the cis gene. In Aim 3 we will develop tools to examine transcriptome-level information in the individual cells of complex tissues and cell populations. These methods will put CTCHGV in position to examine downstream effects of the cis variations we study. The need for Aim 4 developments in support of Aims 1-3 has already been noted, and Aim 4 projects will have wide applicability to biomedical research. These observations illustrate the high degree of integration and innovation in the CTCHGV proposal.

Numerical Targets and Scope Clarifications: The number of genes (Aim 1), cell types (Aim 2), and transcripts (Aim 3), we will actually analyze will depend on the success of our methods and cannot be predicted with certainty. However, based on our track record of innovating high-throughput methods (see Specific Aims) and our experience with the relevant technologies, we prefer setting ambitious goals that may seem risky vs. more secure and unambitious goals, with the understanding that we will review goals and renegotiate them with NHGRI at the end of year 2 of the Center in the light of our own progress, that of our collaborators, and advances in the field generally. We also feel that only ambitious goals will allow us to adequately evaluate and demonstrate the *scalability* of our methods, and we believe scalability will be essential to follow-on application of our methods outside of the Center. With this general statement in mind, our initial targets for main Aim initiatives are: 1000 genes analyzed for cis causality (Aim 1), 50 genes in three iPS-derived cell types (Aim 2), and 1000 transcripts per single cell (Aim 3, both directed and untargeted sequencing; see Aim 3, section 5.3); for Aim 4, see the end of section 5.4. Final and intermediate goals are discussed at the end of each Aim's Research Design section. The targets described here apply to "main" Aim directions but not necessarily to all sub-Aims, some of which deal with analysis of statistical or representative subsets of genes or variants, or demonstrations of related methods or phenomena. For instance, sub-Aim 1.3's analysis of mechanisms by which cis variants control expression is not intended to be performed for all 1000 targeted genes and variants, but only to a small representative subset of genes and variants.

Finally, having described our goals, here we clarify the scope of our Center by identifying several items that we specifically consider to be *out of our scope*: (a) We emphasize again that the 'causation' that we study is *causation by genetic variants of differential cis gene expression*, not causation of disease or organismal phenotype. (b) Nor do we attempt to systematically identify variants that cause differential expression in *trans*, although Aim 3 will allow us to track some downstream consequences of cis causal variants. (c) While ideally we should like to study actual primary human tissues, we will focus on iPS-derived cell types representing human tissues for the reasons indicated in Background and Significance, section 3.1. However, Aim 3 will feature a comparison between one iPS-derived cell type and primary tissue. Nor will we study effects in tissues in non-human animal models. (d) Although in Aim 1.4 we explore how GWAS can relate to our findings, we ourselves will not perform GWAS nor any large scale population analysis.

5.1: Aim 1: We will develop and demonstrate novel methods that identify and characterize natural cis variations that directly affect transcriptional activity in individual humans based on direct modification and testing of combinations of variants in gene regulatory regions in cell lines, and that can be applied to thousands of genes.

5.1.1: Aim 1.1: We will develop and demonstrate novel, high-efficiency methods to create human cell populations containing combinations of natural variations in gene regulatory regions, focusing on zinc-finger nuclease (ZFN)-mediated recombination of externally generated altered insert libraries, and direct modification of human cells using oligo-based methods.

We will develop two main methods for generation of combinatorially altered cells, both of which will use our oligo-based MAGE technology (see Preliminary Results, 4.4). Our principal strategy will be to use MAGE to alter regulatory regions in human BACs in *E. coli*, after which these altered regions will be re-introduced into human cells. The second will develop and apply a version of MAGE that operates directly in human cells. Both of these strategies will involve development of significant technology that will have wide applicability outside of CTCHGV. While the first will leverage a MAGE technology that works efficiently in *E. coli* (see Preliminary Results, 4.4), it will need to be integrated with substantially improved methods for transferring DNA fragments between human and *E. coli* cells (including entire altered regulatory regions of ~100kb), and for

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

efficiently inducing homologous recombination in human cells, in order to replace native with altered regulatory regions. Here we will focus on improving direct transfer by modified bacteria of BAC fragments to human cells, and on use of zinc-finger nucleases (ZFNs) to induce efficient recombination. For the second, we will adapt our MAGE method so that it works directly and efficiently in human cells (including, eventually, induced Pluripotent Stem cells (iPS); see Aim 5.2), taking into account other oligo-based methods for engineering human cells (see Background and Significance, 3.4, and 5.1.1 (ii) below). We will refer to the first strategy as MAGE-BAC/ZFN and the second as MAGE-human. The MAGE-BAC/ZFN strategy is illustrated in Figure 5.1.1-1. In terms of structure, we divide our research plan into four sections, putting the two MAGE sections first: (i) MAGE-BAC (the MAGE and human/*E. coli* transfer aspects of MAGE-BAC/ZFN; with plans for ZFN improvement in (iii)), (ii) MAGE-human, (iii) improvement of ZFN-mediated recombination, (iv) performance evaluations.

As noted in the Research Design Overview above, depending on circumstances, we will sometimes generate altered cells with specific sets of genotypes or haplotypes, and sometimes generate cell populations with combinatorially randomized genotypes or haplotypes. In general, MAGE-BAC/ZFN will enable all of these (see (i) below), while MAGE-oligo will be better suited to generation of specific or combinatorially randomized genotypes than haplotypes as it will not always be straightforward to target oligos to specific alleles.

5.1.1(i) MAGE-BAC: To create libraries of altered cis loci in regulatory regions of a gene, the regulatory fragments on which MAGE will be applied will first be isolated from our CTCHGV subject cell lines and moved into *E. coli* on BACs. After alteration, parts or wholes of regions altered for specific loci, or entire combinatorial libraries, will be transferred back to the original human cells to form cis-altered populations of cells. We describe the MAGE and transfer phases of this work here, while in 5.1.1(iii) below we describe how we will induce recombination of the re-introduced varied regions in the human cells.

5.1.1(i.a) Isolation of Specific Upstream Cis elements via TAR Cloning. We will clone the targeted regulatory region genomic DNA fragments from the subject cell lines onto shuttle YAC-BAC vectors using current methods of Transformation-Associated Recombination (TAR) cloning (88-90). Recent studies have shown that TAR cloning has been used successfully to isolate specific human DNA onto yeast artificial chromosomes (YACs) from human and mouse cell lines (87, 88, 91). Building on these methods (89), we plan to selectively clone all target cis elements by using shuttle YAC-BAC vectors with a 5' targeting-sequence (hook) and a common repeat (e.g., Alu) as a second targeting sequence. Thus, a library of all target cis elements will be constructed at the end of the TAR cloning process. Importantly, our YAC vectors generated by *in vivo* recombination in yeast will contain the F-factor origin of replication, permitting their propagation as BACs in *E. coli*.

5.1.1(i.b) MAGE-generated altered BAC Libraries. Having isolated cis elements on BACs via TAR cloning in *E. coli*, we will use our recently developed automated genome engineering methods ((184) and Preliminary Results, 4.4) to create alterations. Using MAGE, specific single-stranded DNA oligonucleotides (oligos), or pools of oligos, are introduced into an *E. coli* strain which initially contain an isogenic cis fragment derived from the human cell line, and this step may be performed repeatedly on *E. coli* strains derived from previous steps to eventually obtain an *E. coli* strain that contains cis elements with all the desired modifications. We will design oligos that specifically mutate targeted loci within the cis elements contained on the BACs. If we wish to change only a single locus, we introduce only the oligos corresponding to that locus; a small number of altered *E. coli* clones may need to be generated and assayed to identify one that contains the

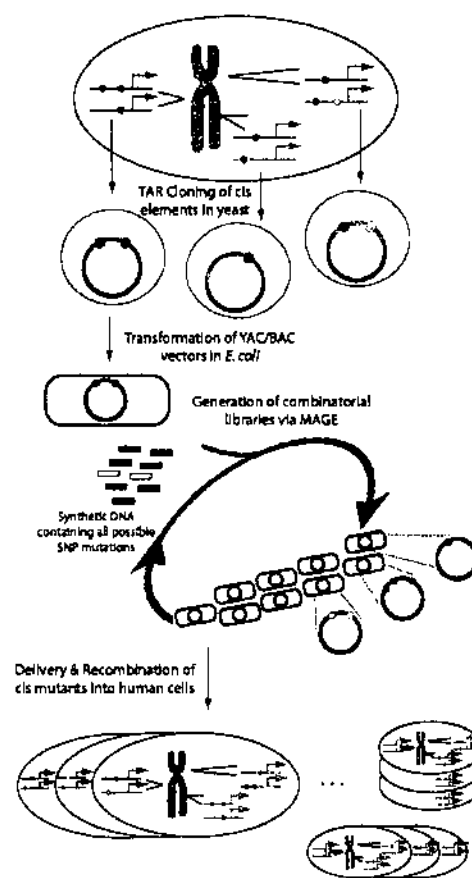


Figure 5.1.1-1: Illustration of MAGE-BAC/ZFN strategy for generating combinatorially modified human cells

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

desired alteration on the desired cis element allele. This strategy can then be serially iterated to make specific multi-locus alterations in the cis element. To simultaneously create combinatorial libraries of each cis element on each cis element allele, all we need do is introduce oligos that modify all loci at once. Upon completion of the MAGE process, the mutated *E. coli* population will have been altered to sample all possible combinations of the targeted cis loci. After transfer and recombination of the altered regions back into human cells, we will assess the performance of our methods as described below in 5.1.2 (iv). This information will allow us to tune the MAGE process so as to improve the efficiency of targeted alteration and the uniformity of combinatorial alterations.

5.1.1(i.c) Delivery of mutated DNA from altered BAC libraries of upstream cis elements into human cells. We plan to optimize the re-introduction of the mutated cis regions generated in (b) into human cells using a number of recently developed delivery and recombination methods: (c.1) Bacteria have been long known to have the capacity to conjugally transfer genomic DNA to other bacteria (171), but more recently bacterial transfer to eukaryotes (58), including mammalian cells (186) has been reported. We will optimize this process for delivery of our altered BAC DNA from *E. coli* to human cells by strategic placement of the *oriT* sequence and selectable markers that flank the region of transfer, factors we have found to be important in our recent work on re-engineering the *E. coli* genome (Preliminary Results, 4.4). (c.2) In a second approach, we will utilize the GET recombination inducible homologous recombination system for the delivery of human genomic BAC clones into mammalian cells (124). The *E. coli* strain DH10B will be used as the host. We plan to introduce four genetic modifications: The λ -prophage from strain DY330 (193) and the *mutS*⁻ gene deletion (33) will enable efficient oligo-mediated λ Red recombination to generate the desired mutations; while *asd*⁻ gene deletion (52, 124) and expression of the *Yersinia pseudotuberculosis* invasin gene will enable DNA transfer. Expression of invasin renders *E. coli* competent to invade HeLa, COS-a and CHO cells by allowing the bacteria to bind to mammalian integrin receptors and trigger their internalization into primary vesicles. Once inside the cells, the *asd*⁻ mutation causes diaminopimelic acid auxotrophy, leading to defective cell wall synthesis and death of the bacteria, making their DNA available to the human host cell (52, 124). (c.3) As an alternative approach to the bacterially-mediated transfer methods above, we will consider use of the HSV family of viruses for the infectious delivery of large BACs. Prior work has shown the successful viral packaging of 150 kb BACs with efficiency of delivery ranging from 25 to 100% into human MRC-5V2 and fibroblast cell lines (105, 183). Importantly, the high-capacity HSV-1 amplicon system permits the rapid transfer of mutated BAC libraries into an appropriate human cell line.

Finally, selectable markers will be used to select human cells to which modified DNA fragments have been delivered. However, these markers cannot be used to ensure integration of the modified fragments because that would require the markers to be integrated as well, and this would confound our objective of changing nothing but the ≤ 5 targeted cis variant loci per gene. Instead we will flank the modified within the modified sequence with the markers so that they will not be integrated during homologous recombination. This strategy has proved successful in our assembly of the *E. coli* genome out of ~145kbps modified fragments described in Preliminary Results, section 4.4.

Potential problems and alternatives: We do not anticipate problems for Tar cloning of regulatory regions (5.1.1(i.a)) as this is a well documented and widely used procedure, nor with performing MAGE on BAC fragments in *E. coli* (5.1.1(i.b)), as we have successfully applied MAGE techniques very effectively (see Preliminary Results 4.4). However, the introduction and homologous recombination of altered regulatory regions into human cells represent new ground and these processes may be inefficient. In that case: (1) It may prove difficult to re-introduce the very large altered BAC fragments intact into human cells. In that case we can break the fragments into smaller pieces and introduce them one at a time. This will likely require a ZFN to be designed for each piece of each fragment. (2) We can use a combination of selectable and counters selectable markers in such a way as to make seamless modifications, a strategy which works well in prokaryotes (104) and which was partially developed for *E. coli* by the Church Lab. This strategy has potential to improve both delivery and integration of large DNA fragments simultaneously. (3) Through our collaboration with the Elledge Lab (see Letter of Support), we can perform genome-wide RNAi and overexpression screens to identify factors that improve the efficiency of human cell delivery of DNA, and then use cell lines modified with the appropriate factors. (4) Finally, note that Aim 4.2 (section 5.4.2(iv)) lays out our plans to develop a segmental genome replacement strategy based on simultaneous use of two ZFNs per regulatory region.

5.1.1(ii) MAGE-human: The second strategy to generate human cells with modified gene regulatory

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

elements will be to develop a MAGE method that works directly in human cells (184). Site-specific gene modification can be achieved by targeting oligos or DNA fragments to the homologous genomic DNA sequence using chimeric RNA-DNA oligonucleotides (RDO), single-stranded oligodeoxynucleotide (ODN), small fragment homologous replacement (SFHR) and triple-helix forming oligonucleotides (TFO) (64). Efficiency of oligo recombination is the key metric in implementing a seamless and selection-free recombination system. This metric, combined with our expertise in achieving highly efficient oligo recombination in *E. coli*, directs our efforts towards similar oligo-directed recombination strategies directly in human cells. In *E. coli*, oligo recombination is mediated by the β protein of the λ Red recombination system, in which the oligo has been proposed to chromosomally integrate at lagging strand synthesis of DNA replication at 25% efficiency (33). Our MAGE technology improved this efficiency to greater than 30% with an ability to introduce multiple modifications simultaneously targeting many chromosomal loci (184). The MAGE technique iterates oligo-based changes through successive populations of cells, and can be used to both generate populations that are 100% modified at particular sites, or to generate combinatorial populations.

Oligonucleotide-mediated recombination has already been used to induce site-specific genetic modifications in select mammalian cell lines, including HEK-293, CHO and embryonic stem (ES) cells (36, 64, 128, 143), with efficiencies of ~0.03-5%. While the mechanisms of oligo recombination in mammalian cells are not known, these studies have revealed important design criteria that we propose to implement and enhance, including: (a) Similar to *E. coli*, the mismatch repair (MMR) pathway negatively affects oligo-directed recombination in human cells (36, 129). (b) Preferentially enhanced gene repair activity of antisense over sense oligos was observed in human cells, suggesting a link to transcription-coupled repair where gene repair by oligos occurs more efficiently when the target gene is actively transcribed (64). Moreover, (c) it has also been shown that chromosomal positioning effects have little or no influence on observed strand bias of oligos and that unmodified (e.g., no phosphorothioate bonds) antisense oligos exhibit 16-45-fold higher rates of modification than sense oligonucleotides (128, 129). (d) The Rad52 protein, mechanistically similar to the single-stranded DNA binding protein β from λ Red, can also be utilized to enhance the recombination of oligos during replication in mammalian cells (64). Given that heterologous ssDNA-binding protein homologs have been shown to function in *E. coli* (34), we also plan to test if these homologs (and others) can enhance oligo-directed recombination in human cells. To complement these efforts, we will also investigate enzymes and factors that are implicated in homologous recombination (HR) and MMR. For example, Rad51, a central enzyme of HR that polymerizes on ssDNA and assembles into helical nucleoprotein filaments, promotes both homology searches in dsDNA and exchange of DNA strands between ssDNA bound with the filament and the homologous dsDNA. Also, the Msh2 protein, a central factor in MMR, is involved in the inhibition of recombination between mismatched sequences, and only upon its deletion can oligos introduce mutation in mouse ES cells (36).

Finally, key to the success of MAGE in *E. coli* was optimization of the electroporation conditions used to deliver the oligos into the cells, and development of automation that iteratively cycled *E. coli* populations through periods of electroporation and recovery, so that modest initial per-cycle modification rates could be amplified significantly. We will develop similar optimization and automation for DNA delivery (see 5.1.1(i.c) above) and human cells.

Potential problems and alternatives: It is possible that oligo-based HR will remain inefficient even after exploring the options above. In that case, (1) we can perform genome-wide RNAi and overexpression screen to identify additional factors that will enhance efficiency with the help of our collaborators in the Elledge Lab (see Letters of Support). (2) We will also explore whether ZFNs can improve oligo-based HR. The latter has been reported to be enhanced by the presence of double stranded DNA breaks induced by I-SceI meganuclease (143). Use of ZFNs with oligos would represent an HR strategy that would allow us to eliminate the Tar cloning and MAGE-BAC components above. However, if oligo-based HR remains intractable, we will pursue MAGE-BAC/ZFN as our exclusive strategy.

5.1.1 (iii) Zinc-finger nuclease (ZFN) mediated recombination: Engineered ZFNs present an attractive direction for recombining the sequences of MAGE-BAC-generated constructs and libraries into the regulatory regions of Aim 1 target genes. ZFNs function as dimers with each monomer composed of an engineered zinc finger array (typically consisting of three or four fingers) fused to a non-specific cleavage domain from the FokI endonuclease (Figure 5.1.1-2a). The zinc finger arrays in ZFNs can be engineered to bind target DNA sequences of interest. Each individual zinc finger binds a 3 bp "subsite" and therefore a ZFN dimer can in principle recognize 18 or 24 bp target sites, depending on the number of fingers in each ZFN monomer. Each

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

ZFN monomer binds to a DNA "half-site" in the full target sequence and introduces a double-strand DNA break (DSB) in a "spacer" sequence between the half-sites. Repair of a ZFN-induced DSB by homologous recombination (HR) with an appropriately designed exogenous "donor template" can be used to introduce a

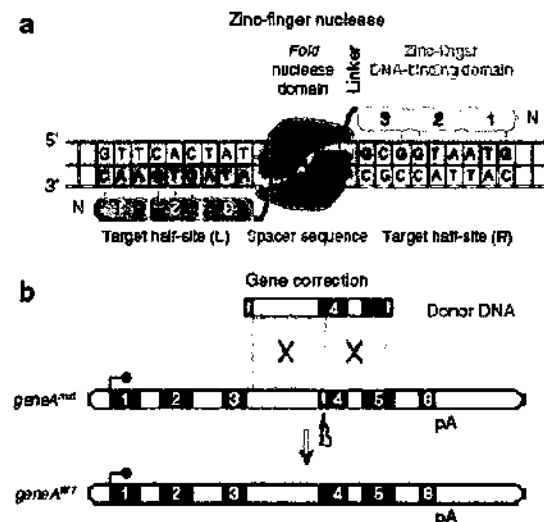


Figure 5.1.1-2. Engineered zinc-finger nucleases (ZFNs). (a) Architecture and application of ZFNs. A ZFN designed to create a DNA double-strand break (DSB) in the target locus comprises two monomer subunits. Each subunit contains three zinc-fingers (1-2-3), which recognize 9 base pairs within the full target site, and the FokI endonuclease domain (green). Dimerization activates the nuclease, cutting the DNA in the spacer sequence separating the target half-sites (L) and (R). ZFN subunits comprising four zinc-fingers that recognize 12 base pairs have also been developed. (b) ZFN-mediated gene disruption and correction by homologous recombination (HR). A DSB (yellow flash) is introduced by the ZFN into mutant allele A^{mut} of a gene. The presence of donor wild-type DNA drives DSB repair through HR vs. error-prone non-homologous end joining, yielding a functional wild-type allele gene^{A^{WT}}. Rather than repair genes, CTCHGV will use ZFNs to mutate cis gene regulatory regions in order to determine their causal role in allele-specific expression. Figure adapted from Cathomen, 2008 (see References).

specific mutation or insertion with very high efficiency near the break (Figure 5.1.1-2b). This method, known as ZFN-induced *gene targeting*, has been used successfully to alter endogenous genes in human cells with absolute efficiencies ranging from 1-50% (107, 118, 178). However, these performance levels are not consistent across all genes and cell lines tested to date, and many additional avenues for increasing efficiency remain to be explored. We will pursue several strategies for improvement that are close at hand in the context of this Aim (1.1). We note that longer-term improvements employing more advanced technologies will also be developed in Aim 4.2 (section 5.4.2 below).

One of the current challenges of targeted genome modifications in human cells is the low rate of native homologous recombination. Although this is greatly improved with ZFNs, error-prone non-homologous end joining (NHEJ) is still a relevant competing pathway of DNA repair which can introduce unwanted insertions and deletions at the break site. We will pursue multiple strategies to enhance homologous recombination and to minimize undesired NHEJ:

5.1.1(iii.a) siRNAs and overexpression cDNAs. The Elledge Lab at Harvard Medical School has already performed a screen which has identified siRNAs that can enhance double-strand DNA break-induced HR of a GFP reporter gene. With them (see Letters of Support), we will test whether siRNAs discovered in this screen can also improve ZFN-induced HR at endogenous human genes. The Joung Lab has validated ZFN pairs that can induce targeted, highly efficient HR events at the endogenous human *VEGF-A* (107) and *PIG-A* genes and also has plasmids encoding ZFNs targeted to the endogenous human *IL2R γ* gene (178). siRNAs from the Elledge Lab screen will be tested at these three loci in a variety of different cell types including iPS cells. Cells will first be transfected with siRNAs and then will be transfected three days later with ZFN-encoding plasmids and donor template DNAs which introduce a restriction site. Four days post-transfection of the ZFN-encoding and donor template DNAs, the frequency of HR in the absence and presence of various siRNAs will be determined using quantitative restriction digest/limited-cycle PCR assays as previously described by the Joung lab (107). All PCR-based results will be confirmed by quantitative Southern blot assay as previously described (107). A screen similar to the one performed by the Elledge Lab could also be performed using cDNA overexpression libraries instead of collections of siRNAs. cDNAs identified by this approach could also be tested using the endogenous human gene HR assays described above.

In addition, we can also design a screen which identifies siRNAs or cDNA overexpression clones that inhibit mutagenic NHEJ. To set up this screen, we will construct a cell line which stably expresses a luciferase gene from a single integrated construct. In addition, we will use OPEN selections to engineer ZFN pairs targeted to sequences in the first quarter of the luciferase gene. To validate these ZFNs, we would demonstrate that they can be used to inactivate the luciferase gene in the cell line we create. These reagents could then be used to screen for either siRNA or cDNA clones that inhibit NHEJ. Specifically, we would first

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

transfect the luciferase-expressing cell line with siRNA or cDNA clone libraries and then introduce plasmids encoding the luciferase-specific ZFN pair. If an siRNA or cDNA clone inhibits NHEJ in the cell, this should result in less inactivation of the luciferase gene and therefore greater luciferase activity.

All factors identified from these screens would effectively transiently reprogram cells to become more "recombination competent", thereby increasing the efficiency of the targeted genome modifications and improving the specificity by limiting unwanted NHEJ-associated mutations.

5.1.1(iii.b) *Small molecules.* HR is most active in S-G2 and cell cycle arrest with vinblastine has enabled the highest rates of HR reported to date (107, 178). However, this drug also induces over 95% cell death when used in conjunction with ZFNs (M. Maeder et al., unpublished data). We will explore whether other agents that induce cell cycle arrest such as indirubin or hydroxyurea can also induce higher levels of ZFN-enhanced HR without the higher toxicity observed with vinblastine. In addition, we will also test whether small molecular inhibitors of NHEJ-specific components (e.g., the DNA-PK inhibitor NU7441) might reduce the mutation frequency due to NHEJ and enhance HR events. We will test the effects of these various small molecules using ZFNs and donor templates which introduce restriction sites at the endogenous human *VEGF-A* (107), *PIG-A*, and *IL2R γ* (178) genes. Initially, we will use a restriction digest/limited-cycle PCR assay (previously validated by the Joung Lab) to quantify HR frequencies at these loci. For those compounds that show increased HR, we will also use limited-cycle PCR/sequencing assays to assess both HR and mutagenic NHEJ frequencies simultaneously. Our expectation is to identify compounds that increase HR and diminish mutagenic NHEJ. If these targeted approaches do not yield results, we can also perform unbiased screens of small molecule libraries to identify compounds that enhance ZFN-induced HR or diminish ZFN-induced NHEJ in cell-based screens similar to those described above for siRNAs or cDNA libraries. In the unlikely event that these screens are non-productive, we will explore inhibitor-free methods that use FACS to synchronize cell populations in S-G2 phase prior to introduction of ZFNs and donor templates.

5.1.1(iii.c) *Longer donor DNA templates produced with MAGE.* To date, ZFN-induced HR performed by the Joung Lab and others have used donor templates with relatively short homology arms (typically ~1.5 kb of total sequence). Studies of gene targeting in mouse and other organisms have used donor templates with significantly longer homology arms. Our expectation is that the use of donors with longer arms should lead to increased frequencies of ZFN-induced HR. We will explore whether 100kbps-sized donor DNA from YACs or BACs (introduced via conjugation) can improve the efficiency of HR. These longer donor DNAs will be made using MAGE technology (see section (i) above).

Potential problems and alternatives: Based on our extensive experience with ZFNs we expect considerable success using them to induce HR of altered gene regulatory regions. If the above steps do not lead to sufficient consistent improvements in HR efficiency, we will pursue the following additional methods (recalling that in Aim 4.2 we will also be exploring more advanced methods for achieving improvements): (d) ***Tethering of donor DNA templates.*** To stimulate faster kinetics of repair, we will create protein-based tethers comprising two zinc finger domains - one that binds the endogenous targeted allele and one that binds the donor DNA template. This approach may both enhance HR and permit use of lower levels of donor DNA, thereby potentially reducing the frequency of random integration of the donor template. (e) ***Recombination "hot spots".*** At present, it remains unknown whether chromatin state can affect the ability of ZFNs to enhance HR. Experiments at the human *HoxB13* gene performed by the Joung Lab suggest that an expressed, open-chromatin state may positively influence the efficiency of HR (107). To explore the hypothesis suggested by this observation, we will design ZFNs that target regions adjacent to DNaseI hypersensitive sites identified from genome-wide surveys. Our expectation is that such sites might serve as potential recombination hot spots for ZFN-induced HR. An understanding of the relationship between chromatin state and HR frequency will enable better choice of target selection to maximize gene targeting efficiencies.

5.1.1 (iv) *Performance evaluations:* For each method we develop above, we will measure recombination efficiencies and also recombination biases using subsets of samples and cell lines generated for a number of genes. It will be especially important to measure these for combinatorial cell populations, as knowledge of efficiencies and biases will be important for accurate simulations and statistical analysis that will guide our isolation of sets of clonally altered cells (e.g., see Figure 5-2 above), as well as the single cell genotyping / ASE assay we will develop in Aim 1.2 (section 5.1.2 below). When analyzing populations of cells in which multiple loci have been modified, we will use our published single molecule gel-polony method for long-range haplotyping (201), which will allow us to identify and phase modified loci for the cis regulatory region and coding regions up to the indicator SNPs. This method is ideally suited to giving us complete

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

information on the distribution of modifications across the regions.

For MAGE-BAC/ZFN, we will characterize the relative rate of random insertions and off-target modifications vs. targeted regulatory region replacements. Because our methods in section 5.1.2 look at many independently altered cells, random insertions and off-target events are not expected to have significant impact on our ability to analyze causality unless they occur with high incidence or operate with high bias on specific regions of the genome. Random insertions can be gauged easily by performing *in-situ* hybridizations with labeled probes from the regulatory regions: multiple integration events should be easily detectable as cells with more than two spots representing instances of the regions. We will characterize off-target events in a representative set of ZFNs by the following method: (a) For each ZFN, we will construct an inactivated version of ZFN (iZFN) that contains the same zinc finger arrays as the active one, but where these arrays are joined to a version of *FokI* whose catalytic domain is inactive. Each iZFN should bind the same locations as its corresponding active ZFN (aZFN) but will not cut the DNA at these locations. (b) We will introduce each iZFN into one or more CTCHGV cell lines and then use Chip-Seq (74, 82, 179) to identify all locations in the genome to which the iZFN binds and to estimate the percent occupancy of the iZFN at these locations. (c) After identifying all target locations with significant iZFN binding, we will design padlock probes that target these locations using procedures in use in the Church Lab (see Preliminary Results, 4.1). (d) Finally, we will introduce the corresponding aZFNs into the CTCHGV cell lines used in (b) and perform targeted sequencing using the padlock probes designed in (c) to obtain actual genomic DNA sequences of locations likely occupied by the aZFN. We will examine these sequences for evidence of non-homologous end joining (NHEJ) events (see Preliminary Results, 4.5, and Figure 4.5-2) and to estimate the frequency of these events at each location. To estimate the frequency of off-target recombinations and NHEJ events relative to on-target recombinations, we will modify (d) by introducing template DNA into the cells along with the aZFN, where the template has a shorter region of homology with the aZFN's genomic target than the genomic sequences targeted by the padlock probe arms. This will ensure that, during the padlock probe capture reactions (Preliminary Results, 4.1), the probes will copy genomic DNA that contains junctions between the template and genomic site of template integration. These junctions will uniquely specify the locations of all template integrations. Note also that the Joung Lab Pending Support and has already obtained preliminary results for steps (a) and (b) (see Preliminary Results, section 4.5).

Potential problems and alternatives: Based on our experience with and Preliminary Results for the methods indicated, we do not anticipate significant problems gathering the performance data described above.

5.1.2: Aim 1.2: We will demonstrate the identification of specific sets of variations that affect cis gene transcription by engineering many combinations of variations and directly observing their effects on transcription, and also by novel methods of assaying complex populations of combinatorially modified cells at a single-cell level.

As described in the Research Design Overview, the basic idea for identifying which cis loci differentially affect transcription is simply to examine transcription levels of cis genes or alleles among cell lines with altered cis variants: those loci for which changes in the cis alleles have no effect can be eliminated from consideration, while those which do have an effect have causal relevance and can be analyzed for interactions with other loci. As noted in the Overview and in Figure 5-1, we will develop two methods for performing this analysis, a simple one based on clonal altered cell populations (the left branch in Figure 5-1), and a potentially more efficient and scalable method that analyzes entire combinatorial cell populations at once at a single cell level (the right branch). We describe these in turn in (ii) and (iii), after describing initial analyses we will perform to identify the sets of genes and cis variants we will consider for the course of the project.

5.1.2(i) Initial identification of genes and cis variations. The purpose of this component is to identify the genes and regulatory regions to be analyzed in all Aim 1.1-1.2 work and to develop associated resources. As already noted (see Research Design introduction), we will use pre-existing, publicly available tissues and cell lines with potential to be transformed into iPS from HapMap, the PGP, the Framingham Heart Study, or other sources. We will give preference to samples for which comprehensive genome sequence is available, particularly if the sequence is diploid. If diploid sequences are unavailable, we will obtain diploid sequences of a large number of genes and their regulatory regions, either by applying the long-range haplotyping methods described in Preliminary Results 4.2, or through our collaborator Complete Genomics, Inc (see Letter of Support). It will be important to use *clonal* populations of cells from these subjects as our methods depend extensively on analysis of differences in ASE, and different clones of the same cells have been observed to be

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

subject to high levels of random allele inactivation (50). The replicates we use should be separate cell lines cloned from the same subject so that they likely represent different clones. In Aim 1.3 we will compare these replicates to assess the impact of random allele inactivation may have on our results, generally. From the diploid genome sequences it will be straightforward to identify indicator alleles for all genes with heterozygous coding regions, and to identify all SNPs and other non-reference sequences in regulatory regions. Our selection of regulatory regions will be from among 100kbps regions upstream of the transcription start site, introns plus segments of adjacent exons, 100kbps downstream regions of documented transcription termination sites, or documented enhancer regions cis to but outside of these distance bounds. From the ~100 variants that might be expected by chance to be in any 100kb regulatory regions, we will use available information (e.g., from the UCSC Genome Browser (81, 93)) on known transcription factor binding sites, conservation, GWAS associations, and other data, to prioritize variations and pick ~5 or less for follow-up analysis. In selecting genes, regulatory regions, and indicator alleles, we will attempt to leverage resources already developed in (199) where possible. We will likely pick one regulatory region for most genes, and will prioritize genes and regions that are known to be associated with human diseases or traits (taking into account cell types and GWAS that will be considered in Aim 2.2, section 5.2.2), which have little repetitive sequence, and for which unique priming sites with good hybridization properties can be designed for cis variants and indicator alleles.

Potential problems and alternatives: The only potential problem we can foresee is if delays arise in obtaining *diploid* sequences for our subject cell lines. In that case we would develop a provisional prioritized list of genes and variations based on imputed haplotypes, and refine this list as diploid sequences become available. As we have only targeted analysis of 50 genes by the end of year 2 (see Intermediate Goals below), there will be ample time to refine the list. Also, in 5.1.2(ii) below, it will be seen that considerable analysis can be done with knowledge only of genotypes vs. haplotypes of variant loci in regulatory regions.

5.1.2 (ii) *Assay via clonal altered cell populations.* The key elements of this strategy have already been described in the Research Design Overview; here we give only some additional details. The first step is to develop, for the gene at hand, a set of altered cell populations that are clonal for a sufficient set of combinations of cis locus modifications to be able to identify causation and interaction. If cell lines are altered for individual cis loci one at a time, one of the strategies described in Aim 1.1 (section 5.1.1), clonal populations for each combination will either be the direct outcomes of the modification procedure or will be easily created from them. For instance, if we use MAGE-BAC/ZFN to make a single variant **A|a** locus homozygous **AA**, the resulting population will be clonal, while if we use MAGE-human and supply **A** and **a** oligos together, we will get all four possible cis regulatory alleles: **A** and **a** on the allele with indicator SNP **x**, and similarly for indicator SNP **y**. Clonal populations for each set of haplotypes can be easily isolated. However, if instead of altering loci one at a time, we generate a population combinatorially modified for all loci, Figure 5-2 above and the surrounding discussion show that by isolating and growing out modest numbers of single cells (~60) from the population, we can get a sufficient set of populations to identify simple causality with near statistical certainty.

The great advantage of the clonal altered cell population strategy is the simplicity of the assays performed on each clonal population. All that is required is that the cis regulatory genotype of each clonal population be identified along with the expression levels of the cis gene. Given that at most five loci will be modified in any gene regulatory region, the genotypes can be identified by five allele-specific genomic DNA PCRs. Expression level can be analyzed either at the overall gene level, or at the allele-specific level; we will investigate both options. When considering overall gene expression relative to a locus **A|a**, significant correlation between overall gene expression for **AA** vs **Aa** vs **aa** genotypes identifies the locus as one in which the allele has causative significance for expression. When considering allele-specific expression levels (ASE), the test is to see whether ASE disappears in homozygous **AA** or **aa** genotypes but is present in **Aa** genotypes. In the simulation of Figure 5-2, the correlation strategy exhibited better sensitivity than a t-test comparing ASE levels between homozygous and heterozygous genotypes. However, this outcome may be dependent on assay variance and other factors, so both overall and ASE measures will be considered. Both may be measured by simple RT-PCRs, with ASE requiring PCRs that are specific to the indicator alleles in the coding regions. Notice that all of these statistical tests require knowledge of only cis region genotypes vs. haplotypes. This relieves us from having to do additional assays to learn the phasing of the various modifications.

Potential problems and alternatives: We anticipate no significant problems with the approach above, except that assessment of interactions between loci may require large numbers of isolated single cells. The approach developed in section 5.1.2(ii) below will be our principal effort for dealing with this possibility.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

5.1.2 (ii) *whole population assay*: In Aim 4.3 (section 5.4.3) we describe plans to develop advanced cell-handling capabilities which will enable millions of cells to be arrayed on the surface of a flow cell, where they can be both assayed biochemically and morphologically via image analysis. Our plan is to use this capability to array a population that has been combinatorially modified for all five targeted cis loci and assay each cell individually for genotype and allele-specific expression (ASE). The rationale for this proposal is that, although the single cell assays may exhibit high error individually, this error can be overcome by aggregating measurements for millions of cells.

Specifically, after arraying the cells on the flow cell surface, fixing, and permeabilizing them, we will add reagents and oligos required to intracellularly amplify each of the $n \leq 5$ cis regulatory sites modified for the gene at hand and perform the multiplex amplification for all cells simultaneously. This step is then followed either by probing or small scale in-situ sequencing that identifies the alleles present at each regulatory site. Several methods will be examined and evaluated for sensitivity, sensitivity, complexity, and cost. The simplest method is to introduce $2n$ primers that amplify the n sites *in situ*, and then perform sequential one-base primer extensions using labeled bases, a method that was successfully applied in gel colonies (203). Another class of methods involves allele-specific amplifications within the cells using two allele-specific primers and a common second primer, where the 3' ends of the allele-specific primers correspond to the alleles and a distinct sequence barcode is affixed to the 5' ends of this primer. The sequence barcodes for each allele can then be interrogated by appropriate small scale sequencing as above, or by *in situ* hybridizations with labeled probes. Several forms of allele-specific amplification can be evaluated, including ordinary allele-specific PCR, padlock probes with nested common and allele-specific PCR primer sequences together, or padlock probes followed by rolling circle amplifications. In developing this protocol, consideration will be given to methods that make the DNA accessible, such as proteases, detergents, and, possibly fragmentation of genomic DNA (27). The read outs obtained from this step are used to classify the cell for the presence of an allele. If only one allele is detected, that allele will be considered to be present with copy number 2 and the other allele to be present with copy number 0. If both alleles are detected, copy numbers of 1 will be assumed for each. Duplicated or repetitive sequences will have been filtered out in initial selection of the genes of interest (see (i) above) and copy number variation is considered in section 5.1.3.

At this point, interrogation of the cis regulatory genotypes is complete, and the next step is to assay ASE for the gene transcript. The cells are now treated with DNase to destroy the genomic DNA corresponding to the indicator alleles. Methods akin to those used above are now applied to amplify and interrogate the indicator SNPs present in the transcript coding regions that identify the transcript alleles, except that the amplification must begin with reverse transcription. Because the cis region loci have been randomly re-assigned, a genotype heterozygous in a locus **A**_{*a*} may be present in both haplotypes, such that the **A** allele may be cis to indicator SNP allele **x** in one cell but cis to the other indicator SNP allele **y** in another. Because we are not resolving the haplotypes in this assay, ASE must be measured in a symmetrical fashion such as $\text{abs}(\log(I_x/I_y))$, where I_u represents signal for the transcript with indicator allele **u**.

Three factors will control the performance of this population assay: the efficiency of altered cell generation (α), the standard deviation of the error of single cell ASE measurements (σ), and the probability of single cell genotyping misreading a genotype (δ). We assume that the predominant genotype misreading error will be failure to detect one allele. This type of error is important because it potentially has a large impact on the statistics of comparing ASE differences between genotypes, for a locus should cause ASE only when it is heterozygous, not when it is homozygous, and this error makes heterozygous loci appear homozygous. Table

parameters	P-value	loci under consideration				
		1	2	3	4	5
$\alpha=0.3$	0.01	6710	24906	85122	282274	921708
$\delta=0.2$	0.001	10972	38378	127448	414995	1337470
$\sigma=5$	0.0001	15353	52143	170587	550051	1760090
$\alpha=0.8$	0.01	445	1148	2729	6293	14291
$\delta=0.1$	0.001	727	1767	4080	9240	20715
$\sigma=2.5$	0.0001	1016	2398	5457	12240	27244

Table 5.1.2-1: Numbers of single cells (N) that must be analyzed for ASE and genotype from a combinatorially altered cell population using the single cell assay proposed in Aim 1.2 to identify, with various P-values, that one of up to five modified loci is responsible for ASE, assuming completely random haplotypes among cells that were successfully modified, that only one of five modified loci controls ASE, that all loci are independent, and two sets of performance parameters (see text for discussion and comparison).

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

5.2.1-1 provides an estimate of the numbers of cells that must be evaluated to detect which of up to 5 loci may be causative of ASE with P-values ranging from 0.01 to 0.0001, given two sets of performance parameters: Table 5.2.1-1 (top rows) considers a conservative set of performance targets—an α of 30% that has already been achieved in some cases (see Research Design Overview and section 5.1.1), with high error single cell assessment ($\delta = 20\%$ and $\sigma = 5\times$ maximum normalized ASE), while Table 5.2.1-1 (bottom rows) represents concerted improvement whereby α is improved to 80% while the error rates are cut in half. While the numbers of cells required to detect cis loci interacting according to models considered in Table 5-1 has not been specifically modeled, the low numbers indicated in Table 5.2.1-1 suggest that good statistics for such interactions are indeed achievable.

The computational analysis required by this assay is as follows: Images of single cells arrayed in the flow cell will be acquired and intensities, obtained in each cell for the various cis loci and allele-specific transcript labeled mini-sequencings or in-situ hybridizations, will be calculated as image analysis features using standard methods already in use in the Church Lab (e.g., (7)). Various additional features based on additional stains such as DAPI may be obtained to determine cell integrity, cell cycle state (which could alter ploidy), or flow cell position occupancy by other than a single cell. If images are acquired for cells arrayed in random positions (e.g., for cells arrayed randomly on slides), these other stains will be used to segment the image into cells and to exclude any segment that is not an isolated single cell. The multiple images acquired will be registered, and genotypes and ASE measures will be computed for each cell as described above and in section 5.1 Overview. Evaluation of genotypes and ASE will employ intensity (for genotypes) and intensity ratio (for ASE) thresholds developed from original CTCHGV subject cell lines that are clonal with respect to genotype. Meanwhile, estimations of α and the distributions of haplotypes generated in combinatorially modified populations will come from Aim 1.1 (section 5.1.1 (iv)), which will also provide information on random integrations vs. replacements of native cis regulatory regions. If random integrations are too common, additional steps will be taken to filter out cell segments that have excessive intensities from genotype images, or more than two genotype intensity maxima. Parameters δ and σ will be estimated from images acquired of non-modified original CTCHGV subject cell lines. Simulations using these parameters will be used to estimate the numbers of cells needed to distinguish between different cis interaction models (exemplified in Table 5-1).

Potential problems and alternatives: The intracellular assays present the key challenge. However, given that only a maximum of 6 loci (5 regulatory DNA and one transcript) need to be interrogated per cell, and that methods similar to what we require have already been developed (163, 202), we believe prospects for success are high, especially given that we will be developing more powerful *in situ* intracellular Rolling Circle Amplification methods in Aim 3 (section 5.3.1.2). The simultaneous querying of genotype and coding transcript will be a new element. Here we can explore modifications to the protocols described above that can eliminate possible complications. For instance, suitably designed ZFNs may be used to cut the genomic loci corresponding to indicator SNP alleles vs destroying all DNA via DNase: Then, RT-PCR can be performed on the gene transcripts without incurring contaminating signal from the corresponding genomic sites, and without destruction of the amplicons created from the cis regulatory loci. Finally, image analysis can be exploited at many levels additional to those described above to filter out any cells for which signals representing genotype content and transcript level cannot be interpreted. For instance, if the readout of cis regulatory locus genotype described above based on detection of the presence of each allele proves error prone, we can add additional conditions such as requiring that, where only one allele has been detected in a cell, the intensity must be $\sim 2\times$ the intensity found for that allele in cells where both alleles have been detected.

5.1.3: Aim 1.3: We will assess the extent to which cis variants identified as causing altered transcript expression may operate through alternative mechanisms such as differential expression of RNA isoforms, differential transcript degradation, copy number variations, and epistatic marks.

Cis variants detected as causing allelic expression bias could actually operate through other mechanisms. It is also possible that observed ratio differences in allelic expression only arise in certain biological contexts. To address these issues we propose using a variety of established techniques to check for the contribution of each of these aspects to our data. These will include splice variants, copy number variants, epigenetic context and allele specific epigenetic differences, and possibly other factors. These phenomena could potentially affect both our measures and our conclusions. For instance, regarding measurement, apparent allele-specific expression (ASE) based on padlock capture of indicator SNPs in the two alleles of a gene could actually be caused by differential splicing vs. differential expression whereby one allele contains

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

more isoforms carrying its indicator SNP exon than the other (170) (see also (203)). Regarding conclusions, it could be that our identification of a regulatory region *cis* variant as a cause of differential allelic expression might actually be artifactual and instead, that the construct with which we introduced altered *cis* sequence disrupted normal methylation patterns. Our primary objective is to assay these phenomena in a subset of our samples to quantify the extent to which they may affect our measurements and our identification of causative *cis* variants, not to comprehensively assess their impact. This subset will include original, unaltered CTCHGV subject cell lines and a selection of altered samples cloned from our combinatorially modified populations for a small set of genes. We will consult with the Center for the Epigenetics of Common Human Disease CEGS regarding our investigations of epigenetic impacts.

To assess the extent to which measured ASE may be due to differential splicing, we will use RNA-seq (185) or the Affymetrix Human Exon Array (http://www.affymetrix.com/products_services/arrays/specific_exon.affx#1_1) to detect splicing variants. More targeted and cost effective approaches, such as custom exon arrays, PCR or the padlock probe-based capture (100, 139), could be used for a selected number of genes. If alternative splicing contributes to the apparent ASE, we will expect to observe the exon (or part of the exon) carrying the SNP marker more frequently in the apparently more highly expressed allele than the other, and to observe more alternative splicing junctions that skip the exon in the less highly expressed allele. These experiments performed on altered sample clones will also reveal the extent to which *cis* variants we have identified as causative may actually alter splicing vs. expression. Measurements of ASE may also be perturbed by random allelic inactivation that differs between clones (50). We will assess this by comparing unaltered replicate CTCHGV sample cell lines, which should not represent the same clones.

Copy number variations (CNVs) in which gene and regulatory region alleles may be amplified or deleted may complicate inferences of causality based on statistical models such as illustrated in Tables 5-1 and 5.1.2-1. By contrast, our measures of ASE should be normalized for CNV. We will use comparative genomic hybridization (CGH) arrays (such as <http://www.nimblegen.com/products/cgh/>) or massively parallel sequencing (25) to detect copy number variation in a selected number of samples before and after recombination.

To assess for allelic bias due to differences in allele methylation or histone modification, we propose to detect epigenetic state in an allele specific manner. This can be done with targeted sequencing capturing locations that contain a heterozygous site of variation (e.g., a SNP). To detect DNA methylation, similar to our recently published methods (10, 37, 199), we will target ~200-bp regions in bisulfite-treated DNA containing an altered site, an unaltered variation site (see section 5.1.2(i)), and a CpG site. The target size of ~200-bp is chosen because it is within both the capability of padlock probes (139, 199), and the read length of the current Illumina paired-end sequencing platform. Allele specific differences in histone modification will be detected in a similar manner by applying padlock probes encompassing variation sites to chromatin immunoprecipitated (ChIP) DNA produced with an antibody to the histone modification of interest (e.g., H3K4 methylation or H3 acetylation). Allele-specific detection of ChIP DNA has already been successfully performed in a microarray context (110) and should be readily translated to sequencing-based methods. If only a very few such regions can be targeted in this manner, we will need to apply our haplotyping methods based on sequencing of dilute DNA preparations (see Preliminary Results and section 5.1.2) to bisulfite treated DNA to obtain the allele-specific methylation profiles. The regions of interest can be targeted by tiled padlock probes to reduce overall haplotype sequencing requirements. In our original, unaltered cell lines, these experiments will be informative as to whether initially measured ASE may have been due to different allelic methylation patterns. We can also trap high molecular weight genomic DNA in polyacrylamide gels, perform *in situ* bisulfite conversion and polony amplification, and then identify alleles and quantitate methylation levels using single base extensions, using the method of (203). It is also possible that causal *cis* variants only contribute to differences in expression in certain epigenetic contexts; for example, a relevant transcription factor may only bind in the context of particular histone modifications (53). In our altered sample clones, allele specific epigenetic measurements may identify cases where *cis* variants affect expression level by means of altering local methylation, or cases where recombination or MAGE oligos (see section 5.1.1) used to create the altered cell have locally disrupted epigenetic state in an incidental manner that is unrelated to introduced sequence variants. To distinguish between these alternatives, we must assess whether altered methylation travels with the introduced DNA or the *cis* variant.

Potential problems and alternatives: We anticipate no significant problems as the techniques described are well-established or methods with which CTCHGV investigators have considerable experience.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

5.1.4: Aim 1.4 We will analyze the relationship between our methods and results and those of Genome Wide Association Studies and characterize their complementary insights into the effects of variation.

GWAS and other studies have identified associations between SNPs and gene expression levels (40, 121, 149, 150, 165, 182), and some of these findings will be used to prioritize genes we will analyze for causal regulatory variants (section 5.1.2 (i)). Here we will examine our findings from the GWAS side and ask what it would take for GWAS to be able to identify cis causal regulatory variants that we have identified. This analysis will clarify the sensitivities, specificities, and amounts of effort required for GWAS vs. the engineering methods developed here to discover and characterize cis regulatory variants controlling gene expression. Once CTCHGV has discovered a set of causative cis variants, we will attempt to estimate the population frequency of the variant, its haplotype block, and its effect size. For any variant that happens to be assayed on platforms designed for GWAS, its frequency should be easily assessed from available GWAS data, but most variants will likely not have been assayed. For these, we will examine HapMap samples. If sufficient sequence data are available (67) we will estimate the allele frequency from the sequences; otherwise we will measure the frequency from HapMap cell lines. To estimate effect size we must consider the tissue from which the variant was identified in the CTCHGV subject. If matching tissue data is available from genotyped samples from the CTCHGV subject's population (125), we will use it to estimate effect sizes and variances. Otherwise we will approximate the effect size based on the degree of differential expression by which the variant was identified by CTCHGV and apply available information to estimate variances (26). Finally, we will consider two models for GWAS. In the common variant model, we will assume that GWAS is performed with tag SNPs from commonly used array platforms. We will characterize the haplotype block containing the variant, identify the tag SNP in greatest linkage disequilibrium (LD) with it, and estimate the population sizes that would be needed to find an association between expression level and the tag SNP given the LD and effect size, using standard statistics and tools (9, 21, 35, 80, 85, 123, 142, 204, 205) for partial and for whole genome searches for cis effects (165). For the rare variant model, we will use the frequency of the variant itself, the effect size, and make comparable computations, this time considering corrections for limited candidate gene sets (126) in addition to partial and whole genome searches for cis effects.

As a second related analysis, we will consider that the expression level change identified for the variant is itself associated with a disorder with one of a fixed range of penetrances, and estimate the population sizes that would be needed to associate the variant with the disorder by the GWAS models above.

Finally, as noted above (Research Design Overview), CTCHGV will not itself conduct GWAS or population studies. However, CTCHGV will communicate with partner Centers and collaborators who do conduct GWAS (such as the Broad Institute) concerning cis variants found to causally impact gene expression level. Our partners will then be able to explore refined hypotheses for the phenotypes studied by the GWAS, and in turn may be able to assess population frequencies for the variant, furthering the analyses above.

Potential problems and alternatives: We anticipate no significant problems with this sub-Aim. It involves applying standard GWAS tools and methods to parameters determined by CTCHGV experiments.

Aim 1 Goals: Final goals: As noted in Research Design Overview, our final goal is the identification of single and/or combinations of natural variations in regulatory region sequence that control differential cis gene expression using the methods described above for 1000 genes. The survey of causal mechanisms in Aim 1.3, and the analysis of relationship between new methods and GWAS in Aim 1.4, will consider representative subsets of genes and variants. Intermediate goals: Again as noted in our Research Design Overview, we will evaluate progress at the end of year 2 of the Center and renegotiate goals as appropriate. We expect we will have processed ~50 genes by that time. Impacts: In addition to increasing biological knowledge about the regulation of any specific genes we have analyzed by identification of causal cis variants, CTCHGV will have developed methods for precise engineering of human cells through ZFN-mediated homologous recombination and oligo-mediated recombination that will have general and impactful application to the understanding of human disease, gene therapy, and personalized medicine.

5.2: Aim 2: We will adapt and extend Aim 1 methods to function in human induced Pluripotent Stem cells (iPS) and then use iPS to characterize the effect of cis regulatory region variations in a variety of derived cell types that represent different human tissues. We will engineer "marked allele" human iPS that are heterozygous in all exons of many genes that will enable analysis of allele-specific

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

transcriptional and splicing effects in diverse cell types.

Overview: The work of Aim 1 depends on the ability to engineer alterations into human cell lines efficiently. Thus, to achieve the goals of Aim 1, CTCHGV will use robust human cell types that tolerate the protocols that implement these engineering steps. To extend these methods to other human cell types via iPS, either these protocols will need to be modified for iPS, or modified Aim 1 cell lines will need to be transformed into iPS. Once CTCHGV develops methods for generating these iPS, we will use them to explore the effects of cis regulatory variants in different derived cell types. The justification for CTCHGV use of iPS-derived cell types vs. primary tissues from humans or other animal model organisms has been provided in Background and Significance 3.3 and the Research Design Overview above. These goals both imply a strong need for automating methods for generation, maintenance, and control of multiple populations of iPS, and efficient methods for differentiating them. This will be the focus of Aim 2.1, while Aim 2.2 will then fulfill the goal of exploring the effects of variations in different cell types. In Aim 2.3, we will use Aim 1 methods to make complex modifications of iPS and develop a potentially highly useful resource for the research community – the “marked allele” iPS. Through this we will both expand and discern the limits of how far one can engineer iPS.

5.2.1: Aim 2.1: We will combine Aim 1 methods with automated techniques for iPS generation and maintenance to enable exploration of iPS with altered cis regulatory regions.

While iPS reprogramming is now widely practiced and becoming more routine, high throughput and rapid generation of iPS cells for large functional studies will require improvements in efficiency and cost reductions. In the present case, we will utilize a retroviral “monovector” that expresses all four factors necessary for efficient iPS reprogramming from one polycistronic expression cassette (77, 153, 189). We are also incorporating small molecules to enhance iPS reprogramming, as reported, thereby enabling even higher reprogramming efficiency from hair, skin, blood (1, 30, 62, 157, 192). We will grow human keratinocytes or fibroblasts directly on 77 micron algae-based microcarriers containing magnetic beads (Global Cell Solutions). These microcarriers are controlled using a magnetic field during media changes, aeration, and stirring. We will develop instrumentation to couple the delivery of viral and small molecules to automated high-density cell culture for iPS reprogramming on the microcarriers. We will use Complex Object Parametric Analysis and Sorting (Union Biometrica) along with Tra 1-60 and Tra 1-81 cell surface markers to identify reprogrammed cells and sort them into microtiter plates. Note that, with reference to Aim 4.3 (section 5.4.3), iPS cells can be made flat for morphology-based selection by eliminating the alginate complex via transient chelation of Ca⁺⁺ and Mg⁺⁺. Importantly, we have observed that hES cells and iPS cells load and proliferate on the alginate complexes without differentiation, and that iPS colonies growing on alginate-based microcarriers can be frozen down without further manipulation. We will develop methods to automate primary cell isolation, iPS cell derivation, and iPS cell freezing and storage. This will enable rapid and affordable distribution of individualized iPS to researchers world-wide. As a proof-of-concept, we will take our original CTCHGV subject samples and a selection of samples modified by Aim 1 and reprogram, derive, and expand iPS cells simultaneously. This will also enable multiplexed exposure of iPS cells to a combinatorial library of differentiation factors (growth factors, genetic factors, small molecules) for directed *in vitro* differentiation and sorting, all directly on microcarriers. By the end of two years, we expect to have a highly efficient, automated platform for generating functional iPS cells and their derivatives, ready for distribution to the research community.

The foregoing will generate iPS with Aim 1 modifications from primary cell samples engineered in Aim 1. We will also attempt to apply Aim 1 techniques directly on iPS developed from original, non-modified, CTCHGV sample cell lines. This will involve testing and optimizing MAGE-BAC/ZFN techniques described in section 5.1.1(i.c) and 5.1.1(iii), and also MAGE-human techniques from 5.1.1(ii), on iPS cell lines.

Potential problems and alternatives: (a) While our collaborators have expanded and cultured hES cells on microcarriers for up to 2 weeks while maintaining pluripotency (see Preliminary Results 4.3), *in vitro* manipulations such as transfection, homologous recombination, and selection may affect their ability to maintain pluripotency. These operations may also result in chromosomal abnormalities. To control for this we will routinely sample for human pluripotency surface markers and karyotype clones, and only propagate those with intact pluripotency and normal karyotype. (b) Currently iPS and hES pluripotency is most effectively checked by 2D visualization under the microscope. A 3D culture system may make it difficult to image pure iPS colonies during culture. If so, we will maintain cells in the 3D system during reprogramming and move to ordinary 2D culture after reprogramming or when high definition imaging is necessary. We and our

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

collaborators will then explore traditional automation and robotics for analysis of the 2D cultures (e.g., using CompacT CellBase at http://www.automationpartnership.com/cb_ibcss/CBsystem_overview.htm).

5.2.2: Aim 2.2: We will differentiate IPS generated in Aim 2.1 into diverse cell types that represent distinct human tissues and characterize the cell type-specific consequences of cis-regulatory variations.

We will proceed on two tracks: (i) We will identify a limited number of genes and cis variant combinations from our Aim 1.1-1.4 set that are implicated in tissue-specific function and, using the methods of Aim 2.1 (section 5.2.1), create IPS populations from subject samples that were engineered in Aim 1 to contain these identified combinations of cis variants for these genes, and observe the effects of variations on cis gene transcription in IPS-derived cell types corresponding to these tissues. In selecting genes and variants for cell type analysis, we will consider the emerging data from large GWAS and ASE studies that identify regions and alleles associated with specific disease phenotypes, with the thought that CTCHGV findings indicating that particular variants cause changes in cis gene allele expression levels may have relevance to research into the corresponding diseases. Cardiovascular diseases, insulin resistance, and obesity are of particular interest not only because numerous associations have been reported (38, 49, 122, 173, 187), but also because methods for *in vitro* differentiation of human embryonic stem cells and IPS into corresponding cell types (cardiomyocytes, endothelial cells, adipocytes, and beta-islet cells) are well characterized. To ensure that CTCHGV can pursue these studies, genes and variants implicated by these studies will be prioritized in the initial selection of genes for which CTCHGV will engineer variations in Aim 1.2 (see section 5.1.1(i)). Once IPS cell populations engineered for combinations of variants for these genes are in hand, we will create clonal isolates of these cells and measure ASE of the corresponding cis genes as in Aim 1.2, and we will do this again after differentiation of the clones into our target cell types. We will compare these ASE profiles with each other and with the profiles already developed from the somatic cell-based populations assayed in Aim 1.2, and we will identify each clone and gene that shows reproducible changes in ASE after differentiation compared to its original unengineered or corresponding engineered original subject cell lines. These changes in ASE for the cis genes will be indicative of upstream regulation that depends on both regulatory region allele and cell type, against the genetic background of our original CTCHGV samples. For these genes we will attempt to identify transcription factors that bind to the engineered sites (170), look for evidence in the literature that they are differentially expressed in corresponding primary tissues, and assay for corresponding changes in expression in our differentiated and undifferentiated cell lines. If a biological effect is documented for the cis genes or the upstream factors, we will test for the biological consequences, including alterations in signaling pathways known to play a role in disease pathophysiology. (ii) Using the single cell ASE / genotyping analysis developed in Aim 1.2, *supplemented with transcriptional assays for tissue-specific expression markers*, we will attempt to simultaneously identify causative cis alleles in multiple cell types developed from the combinatorial IPS populations developed in (i), thus multiplexing Aim 1.2 over both cis variant combinations and cell types.

Potential problems and alternatives: We do not anticipate significant difficulties with (i) as these methods have been generally extensively tested using IPS and human and mammalian embryonic stem cell lines. (ii) If reliable cell type-specific expression markers are available, this will generate few difficulties additional to Aim 1.2(ii) (section 5.1.2(ii)).

5.2.3: Aim 2.3: We will engineer human IPS with "marked alleles" for 10-50 genes and demonstrate their use by characterizing allele-specific transcription and splicing in multiple tissues.

While GWAS/eQTL are currently used to identify genomic loci controlling RNA expression level, the methods of Aims 1.1-1.2 will enable dissection of cis regulatory control down to the nucleotide level, and Aims 2.1-2.2 will make these methods applicable to IPS cells and multiple cell types. Here we will further engineer human IPS cell lines using Aim 1.1-1.2 and develop additional methods that will enable measurement of the effects of sequence variations on RNA transcript structure, function, and cellular phenotype. Finding subtle ASE and expression profile variants with specific changes in allele-specific isoforms and/or function in multicellular environments will be important in itself and will greatly aid in identifying the causal ASE and ultimately the causal nucleotides. As a proof of concept, we propose creating a systematically marked allele-specific genome for a set of 10-50 genes in a subset of our CTCHGV IPS lines. For each gene, an indicator SNP distinguishing each allele will be engineered into a degenerate codon position in each exon where natural SNPs are not already present. Thus, every exon in both transcripts from these trait-associated loci will now be amenable to interrogation in an allele-specific manner. These changes will enable the exon distribution of

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

each transcript allele to be characterized, revealing the presence of allele-specific splicing regulation. These capabilities will enable us to investigate whether allele-specific RNA is cell type-dependent, which could provide many insights into the functional consequences of human variation. In conjunction with Aim 2.2 (section 5.2.2 above), we will incorporate cis variants relevant to cell type into these marked iPS cell lines to assess both ASE and *isoform profile* in cell lines differentiated into particular cell types. We will compare these results with isoform profile information obtained in Aim 1.3 from original CTCHGV sample tissues, to identify cell type specific isoform profile changes in our marked genes. Additionally, we will produce marked allele iPS lines for multiple CTCHGV samples and analyze these for differences in cell type-specific isoform profiles and ASE among individuals.

Potential problems and alternatives: Since "marked allele" cell lines can be generated by successively engineering one gene at a time, there is no problem in principle to achieving this goal using the methods of sections 5.1.1 and 5.2.2. Marked exons may potentially affect mRNA secondary structure and, through this, mRNA processing (92). To limit this possibility we will computationally screen possible "marks" in each exon and implement only those that are predicted to have minimal impact on secondary structure (39, 109, 206).

Aim 2 goals: Final goals As noted in our Research Design Overview, we intend to analyze allele-specific expression using engineered iPS cell lines in 50 genes in three iPS-derived cell types. For sub-Aim 2.3, we will generate iPS lines with marked alleles for a collection of 50 genes in 3 subject cell lines. Intermediate goals Again, as noted in our Research Design Overview, we will evaluate progress at the end of year 2 of the Center and renegotiate goals as appropriate. We expect we will have engineered marked alleles in 5 genes in one subject iPS cell line at that time. Impacts Aside from the direct biological knowledge gained from our analyses of specific genes and subjects, CTCHGV-developed methods for automation of maintenance and differentiation of iPS lines, and for engineering iPS with precise genetic changes, will establish broadly enabling technology for research into disease processes in diverse tissues, and into gene therapy and personalized medicine. We foresee that the creation of useful "marked allele" iPS cell lines for *all* genes (vs. 50 demonstrated in CTCHGV) has high potential to become a research community goal, akin to the creation of yeast deletion strains for every yeast gene.

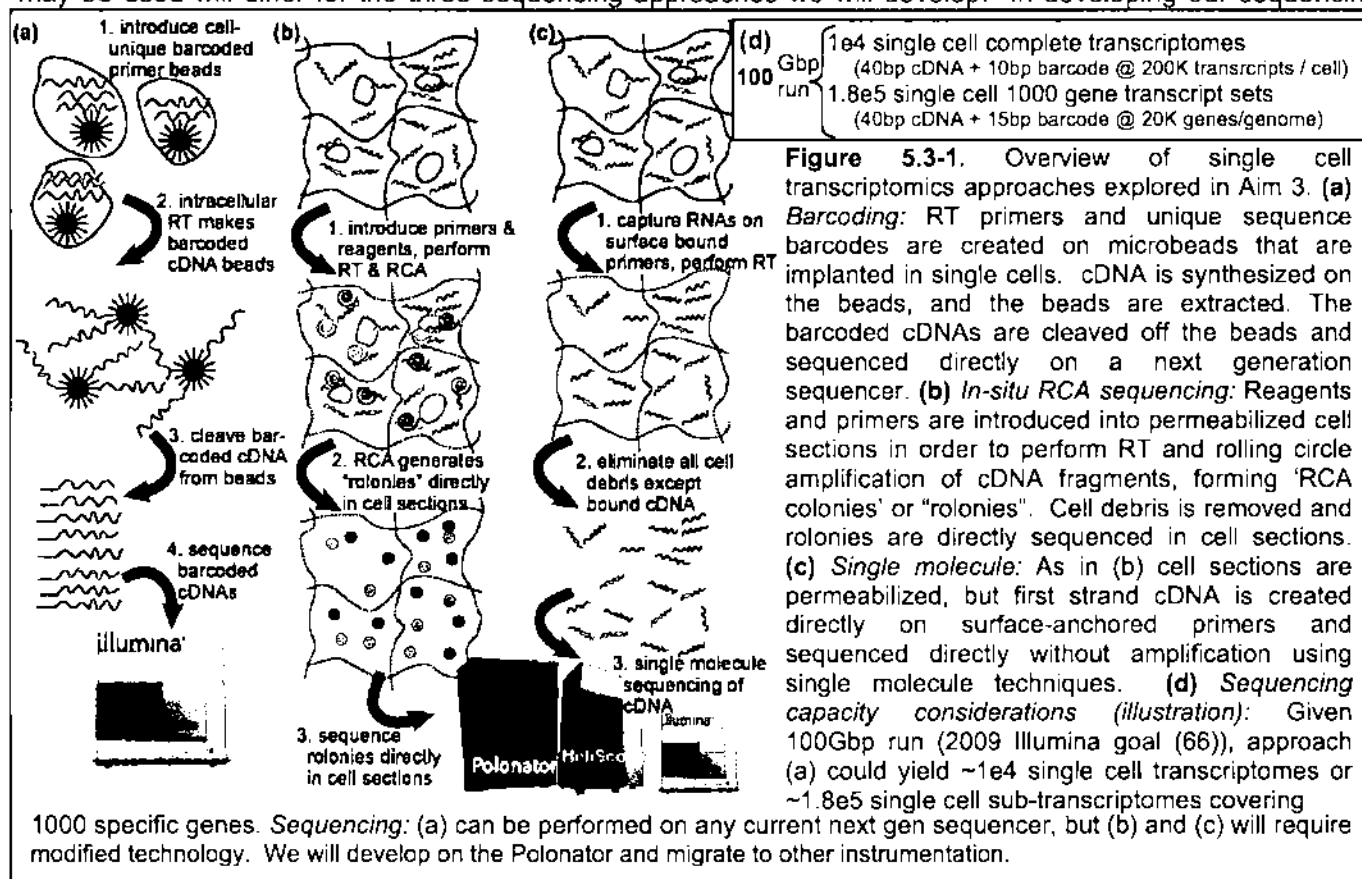
5.3: Aim 3: We will develop novel single-cell in-depth transcriptome assays that are scalable to millions of individual cells in both structured tissues and dispersed cell samples, subject to sequencing capacity. These methods will be used to explore systematic transcriptional effects of genetic variations in different human cell types.

Overview: While single cell technologies are available for gene expression and transcriptome analysis, they are limited by the need to isolate single cells and greatly amplify their minute DNA and RNA content. Laser capture microdissection has greatly improved single cell isolation, and microfluidics improves management of the extremely low concentrations of biological material, but it is still impractical and expensive to isolate and analyze more than a few cells at a time. Here, recent tremendous progress in "next generation" DNA sequencing technology presents significant opportunities, as these technologies have overcome similar limitations by miniaturizing, localizing, and parallelizing operations of DNA capture, DNA amplification, and signal detection. An attractive path forward is to truly integrate DNA sequencing with single cell methods by performing as many of these operations as possible within individual cells rather than in sequencing and preparatory instrumentation. As many different "next generation" methods are available, we will explore multiple approaches for integration. There are also two distinct possible objectives for single cell transcriptomics: (i) *undirected sequencing*, by which one hopes to sequence as many transcripts as possible in every cell in a completely unbiased way, and (ii) *targeted sequencing*, by which one seeks to detect and quantitate large, specific sets of transcripts in every cell, e.g., mRNAs associated with specific biological functions or transcriptional networks. These objectives are complementary, and the choice will depend on the biological problem at hand. We will attempt to develop methods that enable both objectives. The approaches that we will explore are illustrated in Figure 5.3-1, while our strategy is detailed in section 5.3.1. One set of approaches will integrate cDNA synthesis with the introduction of bar codes into cells with single cell resolution, so that the cDNAs bear a sequence tag identifying the cell of origin (see section 5.3.1.1). The cDNAs can be sequenced via normal next-gen sequencing. These bar coding approaches will be useful for undirected sequencing of small numbers of cells or targeted sequencing of up to millions of cells. Where structured tissues are under study, we will investigate methods for associating bar codes with cell location prior

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

to destruction of the tissue. A second approach will develop *in situ* cell sequencing using rolling circle amplification (RCA) of cDNA in either dispersed cells bound to a surface, or in thin tissue sections (see section 5.3.1.2). Finally, to alleviate potential limitations in *in situ* sequencing arising from amplification bias, molecular crowding, and RNA sub-localization, we will explore *in situ* single molecule sequencing (see section 5.3.1.3) by configuring high resolution optical capability into our current Polonator platform (see Preliminary Results, 4.6).

As noted in Background and Significance 3.2, while sequencing capacity is not an inherent limitation to single cell transcriptomes, it may be a practical consideration. Our strategy will be to develop methods by which researchers can control the sizes of the transcriptome subsets that they wish to interrogate, enabling them to use available sequencing capacity to assay large subsets in smaller numbers of cells, or small subsets in larger numbers of cells, as best suits their needs. The technical means for selecting transcriptome subsets will be the choice of the oligos that are used to capture and prime intra-cellular first strand cDNA synthesis. For *undirected sequencing*, the capture oligos will be based on polyT sequences. The 3' degenerate oligo polyT-V (V=A,C,G) will capture all mRNAs and thus yield complete single cell transcriptomes, but smaller transcriptome subsets can be specified simply using more extended and specific 3' ends. For instance, use of polyT-AA, or of polyT-ACG, will yield $\sim 1/12$ and $\sim 1/48$ transcriptomes, respectively. By this means, smaller but still unbiased transcriptome subsets can be obtained from each cell, with the choice of transcriptome size and number of cells left for the researcher to decide based on available sequencing capacity. For *targeted sequencing*, capture oligos will be equimolar mixtures of specific sequences designed to target specific sets of transcripts. For such transcript sets, capture sequences will be chosen based on standard criteria such as uniqueness across transcripts, uniformity of T_m, and secondary structure, with attention to exon boundaries and alternative splicing profiles. The methods whereby capture oligos are created and the oligo mixtures that may be used will differ for the three sequencing approaches we will develop. In developing our sequencing



approaches, we will focus on undirected sequencing first, and then proceed to targeted sequencing. This will allow us to address the common problem of compartmentalizing mRNA capture in single cells first with simple capture oligos before proceeding to more complex mixtures of targeted capture sequences. An illustration of how sequencing capacity might be allocated in different ways is given in Figure 5.3-1d.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Finally, we will develop single cell sequencing approaches for both dispersed single cells and for structured tissues. Here the different methods will differ with respect to how these sample types can be accommodated. We will generally begin testing and development with dispersed cells and move on to structured tissues, where, for convenience, we will do initial work on dispersed cells with human blood cell lines or disaggregated fibroblast cell lines, and initial work with structured tissues using convenient cultured cell lines that have been grown to confluence. As development proceeds, we will switch to using cell types and samples used in Aim 1 (section 5.1) and by Aim 3.2 (section 5.3.2 below).

We expect to develop the three approaches described in Figure 5.3-1 during the first 2 ½ years of our CEGS, and then to determine which to develop further (see Evaluation of the three approaches below). The selected approaches will be applied to biological problems under study in CTCHGV Aim 3 (section 5.3.2): Specifically, we will track transcriptome development of cells undergoing de-differentiation to iPS, or differentiation to distinct cell types from iPS. Single cell resolution is important here because only a small and unpredictable subset of cells achieve iPS de-differentiation, and iPS differentiation, likewise, exhibits a strong stochastic component. Examining individual cells may thus reveal molecular transitions that precede and predict these outcomes that may be hard to observe in any other way. Such observations may lead to new methods for efficient control of de-differentiation and differentiation pathways. Additionally, we will test these methods on structured tissues, comparing in-situ transcriptomes from primary human skin (a complex tissue with many cell types) with iPS cells differentiated into fibroblasts.

Evaluation of the three approaches: We expect all single cell transcriptomics approaches to exhibit trade offs between detection vs. accurate and precise quantitation of transcripts within individual cells. During our development of each approach, we will use common samples and measures that will allow us to compare performance according to these parameters. If a single method has superior performance for both detection and accuracy/precision, it alone will be picked for further development. If no one method is best, or if one works best for undirected sequencing while another works better for targeted sequencing, we may continue development with two methods. To gather the required detection and accuracy data, we will use a common dispersed cell line, and we will assay the transcriptome of this sample as an aggregate population using RNA-seq (84, 185). Using these data, we will define a set of transcripts ("measurement set") consistent with our capture primers that exhibit a range of expression levels expressible as copies/cell, including many with copy numbers of 1 or less. We will then conduct single cell transcriptome assays for both undirected and targeted sequencing on 50-100 cells of this population using each of our three methods. For each transcript in the measurement set, the number of cells in which it appears should approximate a Poisson distribution. We will measure the sensitivity of detection of our methods by assessing the extent to which low copy number transcripts appear as often as they should according to this distribution. We will measure the accuracy and precision of our methods by examining regressions between mean transcript levels across the 50-100 cells and their expression levels measured from the aggregate population, using transcripts in the measurement set with medium to high expression levels. The unexplained variance from these regressions will include contributions from actual stochastic differences between the individual cells and the imprecision of our single cell transcriptome methods. Assuming that actual stochastic differences will be similar for all samples, the method that exhibits the lowest unexplained variance will be the most precise. We will also assay transcriptomes of a number of single cells using microarrays using standard methods (44, 83, 94) and compare the average level of the transcripts observed in these microarrays against the average expression levels seen by our methods. Because both the amplification procedures used to obtain these microarray assays and our own single cell methods may each be subject to systematic biases, we will not use these to judge the fidelity of these approaches but, rather, to assess the presence and degree of any differential biases. We will also perform RNA-seq on individual cells using the technique of (169) and compare with the results of our methods.

5.3.1: Aim 3.1: We will develop and optimize methods that pipeline in-situ single-cell cDNA synthesis to next generation sequencing in ways that preserve cell identity and that can be applied in parallel to 100s to 1000s of cells. We will investigate multiple techniques in support of these methods, including cell bar-coding, in-situ cell sequencing, and single-molecule in-cell sequencing, characterize their performance and limits, and select one for continued development and application.

5.3.1.1 Single cell sequencing via bar coding: Bar coding is chiefly attractive because it can be used with any current next-generation sequencing capability. The main technical issues for bar coding are the generation and placement of unique bar codes in the individual cells to be sequenced. Our initial bar coding

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

approach will use emulsions to encapsulate millions of individual cells with single 1 μm beads displaying approximately one million bar-coded oligonucleotides bearing mRNA capture sequences. This method will be applicable to blood and to disaggregated tissues; a variant of the method applicable to structured tissues will be considered below in (v). Concentrations of beads and cells will be chosen such that there is an average of 1 bead and ≤ 1 cell per compartment. Following emulsion preparation, the cells will

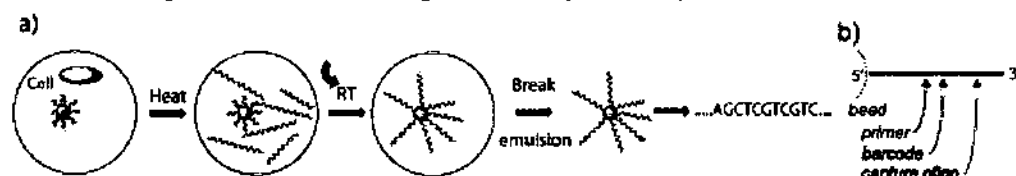


Figure 5.3.1.1-1. single cell mRNA capture and barcoding in emulsion. a) A water droplet containing a single cell and bead. The cell is lysed by heat, bound capture oligos bind to mRNA target sequences, reverse transcriptase is introduced and places the cDNA onto the bead, the emulsion is broken, and the cDNAs collected and sequenced. b) depiction of a bead-bound oligo. 5' primer region (black) allows subsequent amplification of captured cDNA, a bar-code (red) identifies transcripts belonging to the same cell, and the capture oligo (blue) captures the mRNA.

be lysed with heat and the mRNAs will hybridize to the bead-bound oligonucleotides. After mRNAs have been captured, reverse transcriptase (RT) will be introduced to generate first-strand cDNAs, coating the beads with cDNAs (see Figure 5.3.1.1-1). To introduce RT, the beads will either be extracted and re-emulsified in a solution containing RT, or bead-containing droplets will be fused with droplets containing RT. The emulsion will be broken, the beads collected, and the cDNA will be processed and sequenced by RNA-seq or PMAGE (84, 185).

5.3.1.1 (i) *Split pooled DNA synthesis on beads:* To generate populations of 1 micron beads where each bead has a unique bar-code sequence that differs from others in the population, we will employ split-pooled

oligonucleotide synthesis (Figure 5.3.1.1-2) (16). Synthesis will proceed on the bead surface rather than in controlled pore glass, thus ensuring that the oligonucleotides are

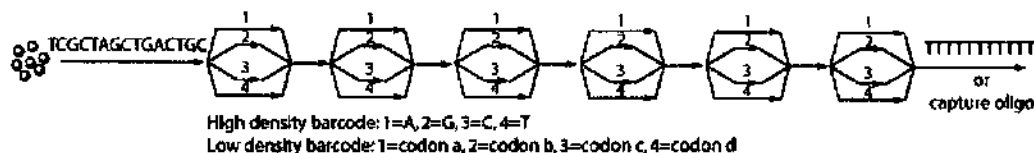


Figure 5.3.1.1-2. Split pooled synthesis of bar-coded oligonucleotides. 5' fixed sequence is grown on the beads. To apply the bar-code, beads are split into four pots and a reagent (single nucleotide or nucleotide triplet) is added to each. The beads are mixed and the process repeated. After addition of the bar-code, beads are pooled and poly-T or capture oligos are added.

displayed on the bead. As bead wetting and swelling properties in organic and aqueous media are important for oligo synthesis and mRNA capture, respectively, we will explore multiple bead surfaces; the swelling properties of polystyrene beads make them ideal for this application, but we will also explore mono-dispersed glass beads, and gold nanoparticles (see below). Oligonucleotides will be synthesized in the 5' to 3' direction on the bead surface, as opposed to the canonical 3' to 5' direction, thus allowing the bead-bound oligo to serve as a primer for reverse transcriptase. As bar-codes must be synthesized error free, we will develop a method for purification of the oligos while on the beads that will result in purities rivaling that of trityl-on RP-HPLC purification. This method exploits the exonuclease resistance of achiral phosphorodithioate linkages. By incorporating this linkage at only the 3' base of the oligonucleotide, lambda exonuclease treatment of the deprotected DNA will degrade incomplete oligos.

We will explore high density and low density barcoding strategies, for readout *via* sequencing and hybridization, respectively. High density bar-codes comprise ordinary sequences to be decoded by standard sequencing and provide 2 bits of information for every base pair in the bar code. Low density bar codes will encode 2 bits of information per 3 base pair "codon" and are decoded by hybridization. Thus, each three nucleotide codon will have 4 variants. To enable 1000 cells to be addressed uniquely with $p < 0.001$ requires high density barcodes of at least 5 bp and low density barcodes of at least 15 bp (five 3bp "codons"). Although low density barcodes are longer, they are not read during sequencing and so their decoding does not contribute to sequencing overhead. By contrast, high density barcodes *must* be sequenced and so yield 5 bp of overhead for each of potentially many millions of sequence reads. Use of low density barcodes is illustrated in Figure 5.3.1.1-3. The hybridization probes used against low density barcodes can be made by combinatorial

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

synthesis of labeled oligonucleotides from triplet phosphoramidites.

5.3.1.1 (ii)

directed and undirected sequencing capture oligos. For undirected sequencing, capture oligos will be polyT with appropriate 3' suffixes as described above. For targeted sequencing, specific capture oligos must be synthesized and then affixed to the beads after

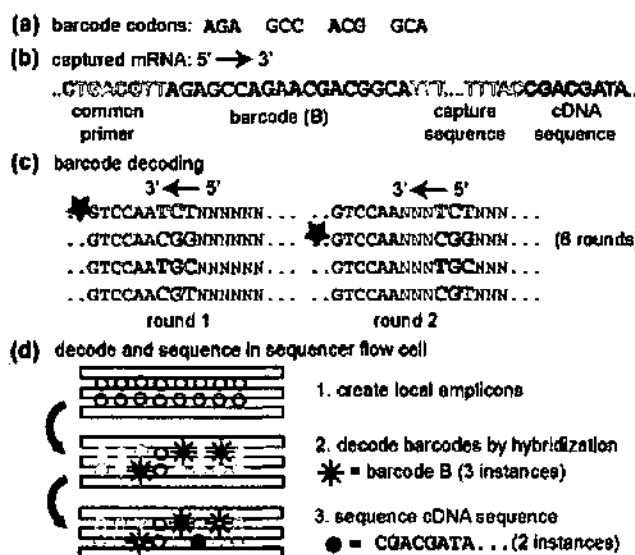


Figure 5.3.1.1-3. Illustration of low density barcodes. (a) Barcode codons. (b) Barcoded cDNA captured from single cell using undirected sequencing capture sequence polyT-AC. All cDNAs with this barcode ("B") come from the same cell. (c) Hybridization rounds used to decode barcodes. Starred primers correspond to barcode B. (d) Assignment of cDNA to cell using barcodes followed by sequencing of cDNAs. Local amplicons of the cDNAs are generated for sequencing. Barcode probing as in (c) assigns amplicons to cells. Sequencing from capture sequence or adaptor identifies cDNA. In this illustration, 3 features are assigned to barcode B, and two cDNAs have sequence of cDNA in (b), one of which is in cell B.

barcode generation. To do this, an equimolar mixture of the capture oligos will be prepared and this mixture will be ligated to the bead sequences using appropriate sets of splint oligos. To allow bead sequences that do not become extended by capture oligos to be degraded, achiral phosphorodithioates at the 3' ends of the capture oligos may be used in a manner similar to (i) above. In developing this approach, we will pay close attention to ligation efficiency and bias, as these factors will limit the size of the target transcript set and the accuracy of quantification. We will optimize these factors initially with commercially available mRNAs that we will mix in different proportions, and then move on to sets of transcripts that are parts of well studied transcriptional networks that are expressed in our samples, with particular attention to networks we expect will be relevant to our intended application in Aim 3.2 (section 5.3.2).

5.3.1.1 (iii) *mRNA capture and barcoding.* Using our emulsion PCR procedure (138), we will place cells and beads in a buffer containing reverse transcriptase, RNase inhibitors, and dNTPs. Emulsions will be formed using the appropriate sets of oils in a single tube by controlled vortexing. Conditions will be optimized such that the average compartment size will be large enough to contain a single cell and bead. The cells will be lysed by heating the emulsion to 95 °C for 10 minutes, after which RT will be added as described above. To capture the mRNA's onto beads we will explore both isothermal reverse and thermocycling transcription. The former is preferred, as it ensures that each mRNA is captured once. However, thermocycling may be necessary to capture all mRNA's in a sample. Following mRNA capture, the emulsion will be broken and the beads collected and pooled for either whole or targeted transcriptome digital analysis.

5.3.1.1 (iv) *Transcriptome digital analysis:* To analyze the captured mRNAs in each cell, the cDNA will be cleaved using a frequent cutter such as *NlaIII* similarly to previous work (84), keeping only the fragments attached to the beads. We will ligate a double stranded adapter to the other extremity of the cDNA, and use it as a common priming site for very limited PCR amplification in pairs with the corresponding common primer synthesized directly upstream of the bar-code. The library will be size selected and sequenced on Illumina GAI next generation platform. Sequencing of one end will convey transcript identity and expression levels. Low density barcodes will be decoded by hybridization (see (i) and Figure 5.3.1.1-3 above), while high density barcodes will be revealed by sequencing from the other end. Transcript abundances can then be obtained simply by counting numbers of sequence features per cell that map to the same gene. For undirected sequencing, transcripts will be mapped to the closest gene whose stop codon is 5' of the sequence read from the transcript in the orientation determined by cDNA capture and preparation protocols.

5.3.1.1 (v) *Structured tissue bar-coding:* To apply these barcoding methods to structured tissues requires a method of delivering the bar-coded capture oligos to cells that have not been disaggregated. Two possible methods are to use ligand-printed surfaces such as those that can be prepared with the use of a device such as the BioForce Nano eNabler (<http://www.bioforcenano.com/index.php?id=295>), or by shooting beads prepared as above into the tissue using a biolistics device. A biolistics approach would only label a

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

small fraction of cells, but the bar codes of these cells could be identified by using the hybridization strategy of (i) above. We will explore these options up to the point of developing an initial proof-of-concept experiment whose execution will depend on whether an appropriate device is available (as these devices have not been budgeted). A simple initial experiment would be to synthesize an array of barcodes on a microarray and attempt to create cDNAs from a permeabilized tissue section that is laid down on the array. As microarray features are typically larger than human cells, this would not support single cell transcriptomics, but could enable transcriptomics of small tissue regions.

Potential problems and alternatives: The greatest potential for difficulty is that coupling efficiency for 5'→3' oligonucleotide synthesis tends to be lower than for canonical 3'→5' synthesis. If these problems prove significant, we will explore polarity switching on solid surfaces, wherein the oligonucleotides are synthesized on the beads in the 3'→5' direction and then inverted *in situ* (96). Another strategy is to synthesize polymer phosphoramidites for all non-variable regions. This is widely practiced with triplet phosphoramidites and decreases the number of couplings required, resulting in higher synthesis fidelity.

5.3.1.2 *In situ* cell sequencing of rolling circle amplified cDNAs: By dint of its covalent linkage of amplified product, rolling circle amplification in various forms has been used to generate extremely compact amplicons primed off of specific genomic and mRNA sequences within individual cells, that can be used for both *in situ* genotyping and digital quantification of transcript abundances (163, 202). However, these applications have only considered very small numbers of loci at a time. Here we propose to greatly increase the scale and generality of these techniques to the point where transcriptome-level information can be obtained from large numbers of individual cells, and we will develop this approach both for dispersed cells and structured tissues. By combining microscopy and image analysis with suitable non-destructive tissue staining prior to *in situ* sequencing, this approach will enable integrated collection of data on cell morphology, protein content and localization, and transcriptome, as well as cell location in structured tissues. We will do our initial development with dispersed cells and move on to structured tissues (see (iii) below).

5.3.1.2 (i) *In situ* library preparation in dispersed cells: Using technology developed for binding cells to a surface described in section 5.4.3, we will capture cells in a dispersed fashion into our flow cell. We will then permeabilize the cells sufficiently to introduce capture primers and reverse transcription reagent, where capture primers are as described above. When specific capture oligos are used, it will be necessary to design them with a common sequence appended 5' to the specific capture sequence to serve as a sequencing primer; for undirected sequencing this is unnecessary as sequencing can be primed off the polyT sequence that is incorporated into each cDNA. Following annealing of the capture primers, we will conduct reverse transcription *in situ*, and then flow in RNase H to degrade the RNA component of the RNA/cDNA hybrids. The cDNAs will then be circularized inside the cells using T4 DNA ligase and short splint oligonucleotides to anneal to both ends of the cDNA, where the splint oligos are designed to hybridize to the common sequence part of the capture oligos at one end while they are degenerate at the other end. Non circularized material will then be digested using exonuclease I and III. The remaining circles will be ready for *in situ* rolling circle amplification (RCA) using phi29 DNA polymerase primed using polyA oligonucleotide. Molecular crowding of rolling circle-amplified mRNA is expected to be a consideration of this approach and will be addressed in the following ways: (a) With dispersed, separated cells, the cells may be lysed to enable localized diffusion of the cDNAs prior to RCA. (b) Capture primers will be redesigned to capture fewer transcripts. (c) Where crowding is not extreme, different sets of sequencing primers may be introduced in the RCA step so that the effects of crowding can be overcome in the sequencing step by initiating sequencing based on one primer at a time (i.e., instead of resolving the crowding on a spatial level, it is resolved by serializing it over time). An issue related to molecular crowding is that localized concentrations of mRNAs may exist in the cell, e.g., in RNA stress granules and P-bodies (5, 6, 47). Very concentrated RNA bodies may be difficult to resolve by these methods; however, detecting and counting them in many individual cells may be an important biological application of *in situ* transcriptome sequencing in its own right.

5.3.1.2 (ii) *In situ* sequencing and digital transcriptomics: Cells will be sequenced using our current single base extension or ligation chemistry on the Polonator platform (Preliminary Results, 4.6) using the common sequence incorporated into specific capture oligos, or polyT for undirected sequencing. Serial sequencing runs on the same cells using different sequencing primers may be required as just described in 5.2.1.2 (i). Transcript abundances can then be obtained by the mapping and counting methods described above in 5.3.1.1 (iv), and tested by application to mixtures of cell lines also described there.

5.3.1.2 (iii) *In situ* analysis of structured tissues: Since *in situ* cDNA library preparation in a tissue

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

section and in dispersed cells (see 5.3.1.2 (i) above) are similar from a technical point of view, we will test our procedures above on a complex tissue to study the various interactions between various cell types and their difference in gene expression. The methodology remains the same with tissue sections, except that we cannot lyse cells as a way of relieving molecular crowding as described in 5.3.1.2 (i). However, with tissue sections there is the option of creating stacks of thin sections of the tissue and sequencing them individually, reconstructing the full cells' transcriptomes by adding together the transcriptomes of the individual layers.

Potential problems and alternatives: It is possible that the splinting strategy for circularizing cDNAs described in 5.3.1.2(i) will not be efficient *in situ*. If so we will focus on padlock probe-based or similar strategies that have been used widely by the Church Lab in other contexts and demonstrated *in situ* at small scales by other groups (163, 202). This approach may complicate capture primer design.

5.3.1.3 Single-molecule *in situ* sequencing: The *in situ* sequencing method of 5.3.1.2 could be improved if the cDNA amplification step of 5.3.1.2 (i) could be avoided, as this would reduce the issue of molecular crowding (5.3.1.2 (i)) as well as reduce the potential for amplification bias. To explore this option, we will test single molecule sequencing of transcripts in cells. A single molecule sequencing instrument, the Heliscope, has been commercialized by Helicos (<http://www.helicosbio.com/>) (55) and another is in development by Pacific Biosciences (<http://www.pacificbiosciences.com/index.php>) (45). Both systems gather sequence information by tracking extensions by single labeled nucleotides of individual transcript molecules that have been captured on a surface, but neither has been developed for single molecule sequencing within cells. However, of the two platforms, the Heliscope is more amenable to this development because it does not require transcripts to be captured within Zero Mode Waveguides arrayed with special geometry on the surface. A Heliscope is available at Harvard. However, because changes in sequencer operation and software will likely be required, we will plan to outfit our Polonator system (see Preliminary Results, 4.6) with optics capable of detecting single molecule signals rather than work on a Heliscope itself. The open-source design of the Polonator will make it much easier to adjust components and software on the Polonator vs. the Heliscope.

5.3.1.3 (i) Biological preparation and sequencing overview: Briefly, our approach will be to layer a thin tissue section or dispersed cells in the flow cell from section 5.4.3 (also used above in section 5.3.1.2 (i)). Initial development will entail sequencing purified single molecules of RNA attached to the flow cell on the Polonator, followed by RNA populations that are anchored by hybridization to polyT-V oligonucleotides. Sequencing reactions will be primed directly off of the polyT-V oligonucleotides and proceed with reverse transcription using single base extension with fluorescent reversible terminator nucleotide technology (see Preliminary Results, 4.6). Targeted sequencing will be tested by capturing RNA molecules on polyT primers that are dideoxy-terminated and subsequent annealing with transcript-specific primers. Each nucleotide will be incorporated directly onto the growing cDNA primed off of the mRNA, incorporated nucleotides will be imaged at every cycle, and the fluorophore and terminator cleaved off to ready the molecule for the next nucleotide incorporation. Once capability is achieved, we will move into attaching dispersed cells on our flow cells, to be followed by 2 micron thin tissue sections as described in section 5.3.1.2. Sequencing of permeabilized cells will be performed similarly to what was previously described in 5.3.1.2. A key issue in sequencing captured transcripts *in situ* is to reduce background that may be caused by cell debris autofluorescence and labeled nucleotide adsorption. To reduce the impact of these factors, we will treat the cells with proteases and detergents after transcript capture to wash away debris, and avoid use of fluorescent labels on nucleotides whose emission wavelengths coincide with cell autofluorescence. Restriction of fluorescent labels on nucleotides will require increasing the number of sequencing cycles, as nucleotides can only be included in the same cycle if they have distinct labels. Heliscope sequencing employs a protocol by which a molecule can be sequenced twice (once in a forward and again in a reverse direction) to reduce sequencing error (55), and we will modify this technique as required to operate in our *in situ* conditions.

5.3.1.3 (ii) Technological requirement of single molecule sequencing: The fundamental requirement for single molecule sequencing is to use optics that enable single molecule resolution. To adapt the Polonator to the technological level necessary, we will redesign our current optical configuration from EPI-Fluorescence to exploit TIRF (Total Internal Reflection Fluorescence). The required hardware modifications needed to retrofit a Polonator for through objective TIRF are: (a) Replacement of the standard 20x Leica objective with 100x high numerical aperture objective, PL APO 100x 1.4NA or similar. The high NA is needed so that the angle of incidence is greater than or equal to the critical angle. (b) Insertion of opaque disk in the illumination path post-collimation pre-camera path so as to only permit fringe light rays from reaching the back aperture of the objective. c) Flowcell skirt addition via adhesive gasket to hold immersion fluid. Outside of the optics, other

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

components such as the optical train, stage motion, fluidics delivery, scanning capture, and algorithms will only need minor tuning to adapt to chemistry in cells. Software changes needed for implementation of single molecule sequencing should be minimal, although this depends on the ability to use reversible terminator nucleotides (vs (55)). If reversible terminator nucleotides are not successful, we will proceed with unterminated nucleotides and make software modifications required to analyze homopolymer additions, as in (55).

Potential problems and alternatives: Detection of base incorporation at a single molecule level in the presence of cell debris will be very challenging technically. If we cannot detect incorporation reliably after efforts to clean up cell debris, we will not pursue this strategy and focus exclusively on the approaches of section 5.3.1(i) and (ii).

5.3.2: Aim 3.2: We will use these single cell transcriptomics capabilities to characterize the transcriptional state differences in cells bearing artificial and natural variant combinations from Aim 1, and from cell types developed from iPS from different genetic backgrounds.

As noted in the Aim 3 overview (section 5.3), our strategy will be to develop and evaluate three single cell transcriptomics approaches in the first half of our CEGS, and proceed to further development in the context of demonstrations of the best approach(es) in the second half. This sub-Aim describes our plans for these demonstrations and their integration with CTCHGV Aims 1 and 2. We will proceed through four series of experiments that start with single cell transcriptome sequencing of dispersed cells and mixtures whose results can be confirmed easily by other means, and proceed to in situ transcriptomics on structured tissues.

5.3.2 (i) Preliminary experiments on cell mixtures We will create mixtures of CTCHGV cell lines that are expected to exhibit transcriptional differences in different fixed proportions, and gauge the extent to which we can observe single cell transcriptomes that bear these differences in approximately the same proportions. These experiments will both test the performance of our approaches, and will also help define optimal identities and sizes of the transcriptome subsets interrogated by our targeted and undirected sequencing methods. Among the mixtures we will consider are: (a) Original, unaltered CTCHGV cell lines corresponding to different tissues (if available). (b) Mixtures of original, unaltered CTCHGV cell lines from different subjects, if in the course of Aim 1 (especially Aim 1.3, section 5.1.3) we have observed transcriptional signatures that differ between samples. (c) An original, unaltered CTCHGV cell line, and the same cell line which has been modified (by techniques of Aim 1.1, section 5.1.1) to contain an integrated GFP gene. (d) The two cell lines from (c) where the GFP-containing cell line has additionally been modified by deletion of both copies of a major transcription factor. Here it will be of interest to see how well presence or absence of GFP correlates with expected changes in the transcriptional network controlled by the transcription factor. (e) An original, unaltered CTCHGV cell line grown in two conditions, e.g., CTCHGV fibroblast cell lines grown in the presence of vs. the absence of serum (70).

5.3.2(ii) Downstream consequences of cis regulatory variations We expect Aim 1 to identify cis variations that control key transcription factors. We will take original CTCHGV cell lines and/or versions of these cell lines altered in Aim 1 to maximize differences in expression of these factors, and first assess clonal outgrowths of these cells for mean expression levels over the entire transcriptome by normal array or RNA sequencing methods. From these data we will then identify a small number of transcripts that exhibit significantly different average expression levels, and assess a large number of cells of each population by in situ hybridizations targeted to these transcripts to characterize the *distribution* of expression levels in individual cells vs. the mean expression levels captured initially. Finally, we will perform single cell transcriptomic sequencing of these cell lines by our Aim 3.1 methods and assess the extent to which transcripts found to differ at a mean level over the population are also found to differ in mean level across the individual cells, and whether the distributions of individual cell transcript levels found within these cells correlates with the distribution found by in situ hybridization. Notice that these experiments will be performed on the individual cell lines separately, not on mixtures as in (i) above.

5.3.2(iii) Differentiation and de-differentiation of iPS Here we integrate our single cell transcriptomics methods with Aim 2. We will take a subset of the CTCHGV iPS cell lines that are being differentiated into cell types representing different tissues in Aim 2.2 (section 5.2.2), extract and preserve aliquots at different time points, and perform single cell transcriptome sequencing on these time point samples. We will also take original (or Aim 1-modified) CTCHGV cell lines that are being de-differentiated into iPS, similarly preserve aliquots at different time points, and perform transcriptome sequencing on these samples. The first set of these experiments should exhibit a progression of subpopulations of cells that ultimately assume

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

transcriptional characteristics of target cell types, but in view of the stochastic nature of differentiation, it may also exhibit subpopulations that do not. A key interest will be to look for any transcriptional characteristics that appear to anticipate the assumption of target cell type identity, as these may give insights into better ways of controlling differentiation. Similar considerations will apply to the de-differentiation experiments.

5.3.2(iv) in situ structured tissue In our main demonstration of in situ single cell transcriptomics of structured tissue, we will perform in situ transcriptome sequencing of primary human skin cells from a CTCHGV subject, and compare the results with single cell transcriptome sequencing of iPS cells from the same subject that have been differentiated so as to yield fibroblasts. As human skin is a very complex structured tissue, we will expect to see a range of distinct transcriptomes in the primary cells, only some of which correspond to the iPS-derived fibroblast transcriptomes. This experiment will be of interest because it will reveal the extent to which single cell transcriptomes vary across a primary sample, how much transcriptomes within a cell type within the sample may vary according to the locations of the cells in the sample, and how much iPS-derived cells of a type within the sample resemble their primary cell counterparts. Our ability to proceed with this experiment will depend on the availability of an appropriate tissue sample (see Research Design Overview and Aim 1.2(i), section 5.1.2(i)).

Potential problems and alternatives: Success on Aim 3.2 above will depend on our success in Aim 3.1, and we have therefore designed the applications above as a series of tests of Aim 3.1 methods that are graded in difficulty. We will take this series as far as we can and use any points of failure to inform further Aim 3.1 work on methods development, and thus move between Aims 3.1 and 3.2 iteratively.

Aim 3 goals: Final goals As noted in our Research Design Overview, we intend to demonstrate single cell transcriptomes (both targeted and undirected sequencing) for 1000 transcripts per cell. This demonstration will be on whichever of the three approaches we deem to be most promising half way through the CTCHGV five year period (see *Evaluation of the three approaches* above). Intermediate goals Again, as noted in our Research Design Overview, we will evaluate progress at the end of year 2 of the Center and renegotiate goals as appropriate. We expect we will have succeeded in interrogating 100 transcripts per single cell at that time by at least one of our approaches. Impacts CTCHGV-developed methods for obtaining single cell transcriptomic data will greatly broaden the ability to understand the distinct roles of the different cell types that participate in complex organisms in their actual, structured tissue contexts. Although single cell transcriptome-level information is obtainable today, current methods are not scalable to large numbers of cells and do not take advantage of the greatly increased throughput of next-generation sequencing. Additionally, CTCHGV's development of both targeted and undirected transcriptome sequencing methods will enable considerable flexibility in application and optimal utilization of sequencing capacity.

5.4: Aim 4: In support of Aims 1-3, we will develop innovative and widely applicable methods for high-throughput synthesis of long DNA constructs, highly efficient homologous recombination in human cells, and highly multiplexed single cell handling that enables sorting based on morphology.

Overview: Central to CTCHGV strategy for Aim 1 is the use of Zinc-Finger Nucleases (ZFNs) to support the engineering of regulatory sequences in human cells. As noted in Background and Significance section 3.4 and Research Design section 5.1.1(iii), the ability to engineer ZFNs targeted to specific genomic sites has matured to the point where both academic research consortia and companies are now generating customized ZFNs (134). CTCHGV will make use of these capabilities with efficiency improvements described in Aim 1.1 (section 5.1.1(iii)), but in support of its broader Aim 1 goal of making these techniques scalable to thousands of genes, here in Aim 4 we will apply our expertise in synthetic biology and zinc finger engineering to two projects (Aims 4.1 and 4.2) that will improve both scalability of synthesis and the targeting range of ZFNs generally. Meanwhile, Aim 4.3 will use elements of the system proposed in Aim 4.1 to improve cell handling that is needed in Aims 1.2 (section 5.1.2(ii)) and Aim 3, in a way that will enable a new form of cell sorting that expands the capabilities of FACS. All three of these projects involve development of highly innovative technology that will have very wide application in biomedical research generally in addition to their supporting roles in CTCHGV.

Aim 4.1 (section 5.4.1) will address scalability and accuracy of synthesis of ZFNs. ZFNs comprise two subunits, each of which is a fusion of three to four tandem Zinc-Finger domains (ZF domains) that enable specific recognition of a DNA sequence with an endonuclease *FokI* (section 5.1.1(iii)). Individual Cys₂-His₂ ZF domains are each about 30 residues long, with specificity mainly conferred by six residues in or adjacent to the

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

domain's α -helix (72, 130). DNA that codes for these 30 residue domains can be synthesized as single DNA oligonucleotides (oligos), so that, in the simplest scenario, to synthesize genes coding for a ZFN subunit targeted to a specific site would require synthesis of three to four site-specific oligos and a small set of splice oligos, followed by enzymatic assembly with common DNA scaffold components that code for the rest of the subunit. Since two subunits are required for a ZFN, synthesis of 1000 specific ZFNs entails ~2000 such operations, each requiring ~10 site-specific oligos (including splice oligos). We (see (174)) and others have recently turned to release and enzymatic assembly of oligos from oligo chips as a low cost method of implementing such synthesis tasks. Currently these methods experience yield and accuracy limitations due to the considerable crosstalk during annealing (and ligase or polymerase) assembly reactions that arises from the release of massive numbers of oligos into small numbers of pools. In the context of ZFN synthesis, we will develop technology for massively parallel hierarchical synthesis that, in stages, sequesters oligos and subsequently assembled DNA fragments that correspond to thousands of individual target DNA constructs into separate compartments, so that assembly of these constructs is not hindered by crosstalk. This technology will innovatively integrate on-chip oligo synthesis and assembly, DNA sequencing, and light-directed release of arrayed fragments.

Aim 4.2 (section 5.4.2) will address the comprehensiveness, specificity, and efficiency of ZFNs. Each ZF domain of a ZFN subunit recognizes 3 base pairs, and specificities for recognition are subtly different for each of the ZF domains linked in tandem in a ZFN subunit. To be able to specifically target any half site in the genome, we need libraries that cover all $4^9=262,144$ possible 9 bp sites. To achieve this, we will use ribosome display methods (196) to sample all possibilities from very large combinatorial libraries engineered to cover this target sequence space.

One of the technological elements in Aim 4.1 is that we will array micron sized features (colony beads) in a flow cell with a light-labile chemical anchor, analyze them (specifically, sequence the DNA on them), and then, based on the results of analysis, release specific sets of beads together by directing light on them. With suitable modifications, these methods can be applied to cells as well as microbeads, and Aim 4.3 will develop these modifications. The cell arraying component will have immediate application in Aim 1.2 (section 5.1.2(ii)), where it will improve image analysis of cells, and also Aim 3.1 (5.3.1.2(i)). Meanwhile, the extra modification of using light-labile chemical anchors for the cells will enable a novel FACS alternative that will allow cells to be sorted not only by markers and general optical properties, but also by morphological features revealed by image analysis. We will demonstrate this capability on an application that integrates other CTCHGV Aims.

Aspects of Aims 4.1-3 will be developed on the Polonator (see Preliminary Results, 4.6), which provides the microscopy and image analysis, sequencing, flow cell control, programming, and open source access and compatibility that allow us to modify and integrate the new elements quickly and easily. However, we will make any modifications open source and use our close collaborations with companies (see Data and Materials Dissemination) to encourage incorporation of our innovations into commercial products.

5.4.1: Aim 4.1: We will develop a platform that integrates DNA synthesis and sequencing and uses sequence information to assure synthesis of DNA constructs with extremely low error rates.

A high level view of one version of the platform we propose is given in Figure 5.4.1-1. As is done today (e.g., (174)), the sequences of the large DNA constructs to be generated are first analyzed to determine a set of construction oligos with appropriate size, overlaps, common primer sequences, and Tms for correct amplification and self-assembly (Figure 5.4.1-1(a)), and these oligos are chemically synthesized on a single stranded DNA oligonucleotide array (as in (174)). The assembly process is illustrated in Figure 5.4.1-1(b). The construction oligos are cleaved from the array and amplified clonally on microbeads using emulsion PCR (155). The microbeads are then arrayed on a flow cell for sequencing, but here the beads are attached to the surface using light-labile linkers so that particular microbeads can be subsequently released by direction of narrow beams of light on them. The beads are then sequenced to simultaneously locate oligos that are part of the same large DNA construct, and to verify that the oligo sequences are error free. For each large construct, light is then directed to its sequence-verified oligo beads so that these can be flowed out of the flow cell and captured into an independent compartment for subsequent multiplex assembly. In place of the microbeads used for illustration in Figure 5.4.1-1, oligos or assembly products could be covalently immobilized (e.g. using EDC/NHS chemistry) and amplified from single molecules using polymerase (or ligase) chain reactions – thermal cycling (PCR) or isothermally (e.g. RCA, hRCA, SDA, HDA, PWGA (<http://www.biopelix.com/technology.asp>)) using zero, one or two immobilized specific or general primers or no

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

primers at all (as in PWGA). The resulting polymerase colonies (polonies) can be sequenced by any of the "next-generation" DNA sequencing chemistries -- e.g. polymerase with FL-dNTPs (117) or ligase with 5-mers to 9-mers (155), MPSS (15), SBH, etc. Instead of release of polonies with correct sequence for subsequent assembly, polonies which have the incorrect sequence could also be selectively destroyed or released -- e.g. via photo-caged nitrobenzyl linkages. For very large constructs, the process is amenable to iteration in that oligos can first be assembled into construct fragments, and the fragments then combined for subsequent assembly. We will do initial development on the Polonator to simplify integration of sequencing and light direction (see 5.4.1(iii) below), but will consult with commercial providers of compatible instrumentation to abet technology transfer.

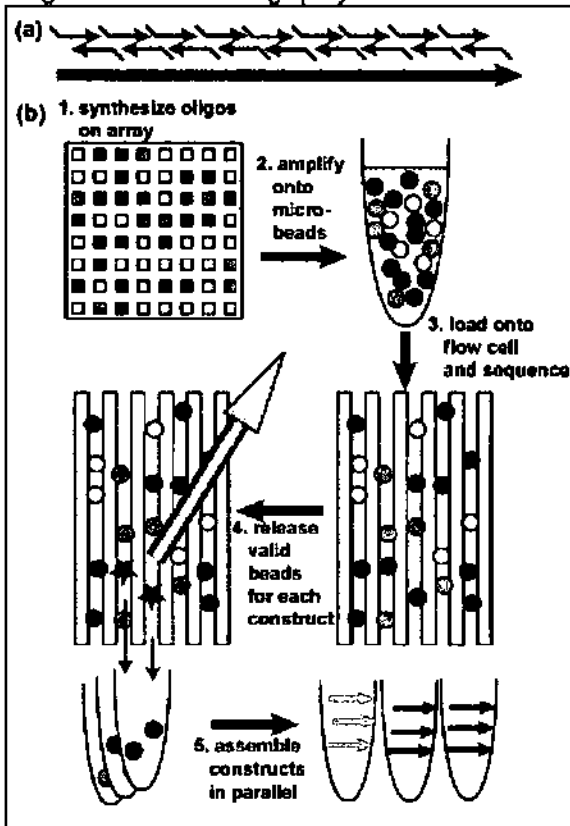


Figure 5.4.1-1. Schematic for one way of integrating DNA sequencing and synthesis for high-throughput reduced-error synthesis of large constructs. (a) Large DNA construct is analyzed into oligos with appropriate overlaps, uniqueness, Tms, as needed. (b) Processing pathway from synthesis of oligos on array for multiple constructs (represented by different colors) to multiplex synthesis. Amplification in (2) is illustrated as emulsion PCR as in (155). Microbeads are loaded onto flow cell using light-labile chemical attachments (see text) and sequenced on the flow cell (3). For each construct, light is directed to microbeads with sequence-validated oligos for the construct for release and capture (4). Assembly of all constructs then proceeds in parallel (5).

These steps overcome crosstalk between masses of oligos by separately collecting and assembling oligos and fragments that are part of the same construct. The sequencing step also overcomes the high rate of error incurred during chemical synthesis of oligos on the array, which, at ~0.5% error per addition (12) can be ~33% for synthesized 80-mers. While current methods, such as mismatch-sensitive hybridization (174), mutS binding (17), MutHSL cleavage near mismatches (160), and direct cleavage at mismatches (11) allow synthesis and assembly errors to be incorporated and then allow them to be filtered out, the method outlined in Figure 5.4.1-1 avoids their incorporation in the first place.

Development of this method will involve small modifications to the Polonator bead and flow-cell preparation protocols to support loading the beads to the flow cell with light-labile chemistry. Towards this end, phosphoramidites containing the photo-labile nitrobenzyl group will be incorporated into the oligonucleotides that are used to "cap" the bead-teathered DNA with the appropriate attachment chemistry. The Polonator may also need minor modification to prevent entry of stray light that could inadvertently release beads. More substantial modifications must be made to support the release of sequences by light direction. Suitable control of light direction can be achieved either by using a Digital Micro-mirror Device (DMD) or a Liquid Crystal Display (LCD). For simplicity, cost, and attainability, we will focus on the DMD approach with initial testing on the Polonator. The array will be used in conjunction with the standard Polonator illuminator and an appropriate photo cleavable (360nm) linker (see Figure 5.4.1-2). For this approach to work, the illumination path must be modified to allow the image of the DMD to be projected on the substrate. The optical path is modified as follows: a) Assemble a Polonator filter block consisting of a standard 50/50 beamsplitter and 360nm excitation filter. The 50/50 beamsplitter is used instead of a 100% mirror facilitating focus and alignment of the DMD to the camera CCD array. b) Place the filter cube in the Polonator filter wheel to allow patterned illumination on the substrate, and c) Insert a tube lens before the 360nm excitation filter allowing the image of the DMD array to be collimated. d) Place the DMD at the focal plane of the tube lens and illuminate the array at an angle with the current 300 watt xenon source. The light reflects off the DMD and is collimated by the tube lens: collimated light then reflects off the 50/50 beam splitter onto the specimen and back up to the camera. The shutter and motion axis allow this selective release to be accomplished over the full area of the

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

substrate.

Use of this technology to generate large numbers of ZFNs required for other CTCHGV Aims After the DNA sequencing and bead release components have been successfully integrated, we will implement additional automation to manage the synthesis of large numbers of ZFN proteins, particularly for Aim 1. We will attach an autosampler which will take beads that are released into the flow-cell volume and feed them into 384 well plates. The same autosampler can then be used as the platform for hierarchical synthesis. The main issue is that the liquid volume after flushing the

flow cell will be about 300 μ L, more than will fit in the well. To accommodate this we can use filtration and apply vacuum while dispensing into the well to remove excess liquid. An alternative would be to use magnetic separation, which would require design of a concentrating chamber. This could be developed using microfluidics. Once automation has been developed, we will design and order the oligonucleotide arrays required for building the ZFNs and proceed with actual synthesis.

Potential problems and alternatives: As the DNA sequencing, DNA synthesis, and DMD technologies are already individually well developed, the main issues that will arise are with integration, and we have laid out our key approaches above. As ZFNs are very simply structured, the problem of synthesis is particularly simple, as construction of each ZFN can be accomplished by the addition of a small set of oligos to standard fragments encoding the rest of the protein. The only other novel component to be integrated is new bead attachment chemistry. Here the main issue is likely to be amount of non-specific absorption to the surface, and we expect we can reduce this easily with different coatings.

5.4.2: Aim 4.2: We will improve zinc-finger nuclease (ZFN)-mediated homologous recombination in human cells by engineering a comprehensive zinc-finger archive, by developing novel methods of delivering ZFNs into cells, and by developing a "segmental genome replacement" strategy.

5.4.2.i: Engineering a comprehensive zinc finger archive: As noted in Preliminary Studies, section 4.5, ZFNs are dimers, the monomers of which each contain tandem arrays of three zinc fingers, which, at full specificity, would be enough to uniquely specify sites in the human genome; however, to date, OPEN zinc finger pools (see Preliminary Studies, 4.5) have been constructed for all three bp subsites of the form 5'GNN at all positions in a three-finger domain (48 pools) and for a smaller number of the 5'TNN subsites (18 pools) ((107) and M. Maeder, J. Foley, & J.K. Joung, unpublished). This limits the targeting range of OPEN to finding potential ZFN sites on average only once every 200 bp, the same range that is available commercially via the CompoZr™ zinc finger engineering service from Sigma (<http://www.editforthebetter.com/FAQs.aspx>). As gene targeting efficiency drops off with increased distance between ZFN-induced double stranded breaks (DSBs) and the desired alteration in mammalian cells (41, 162), improved targeting capability will be needed if ZFNs are to be used widely for homologous recombination (HR)-mediated targeting of gene cis regulatory regions in Aim 1 (section 5.1.1) and many other needs of the research community. We propose to complete the archive of OPEN zinc finger pools and to perform a comprehensive series of selection experiments to identify three-

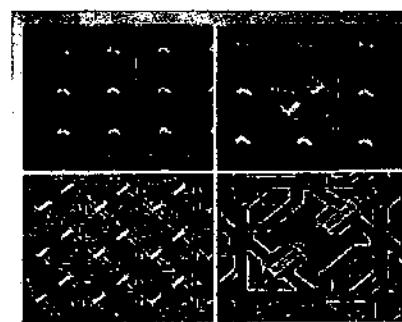
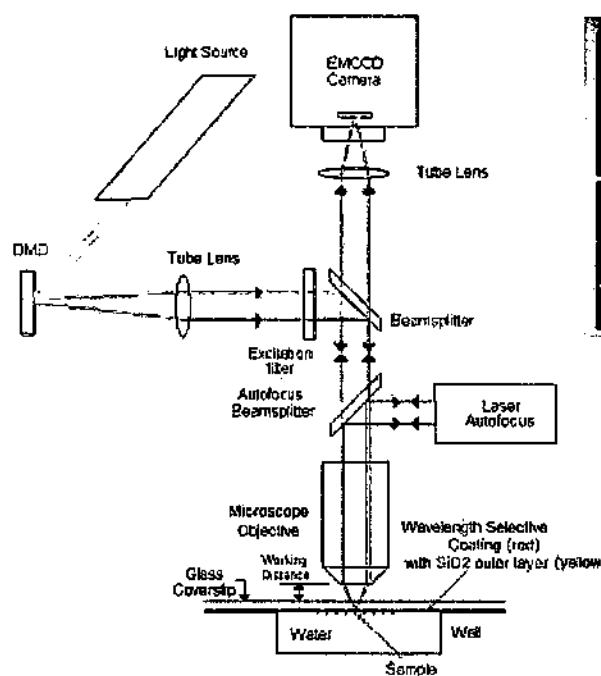
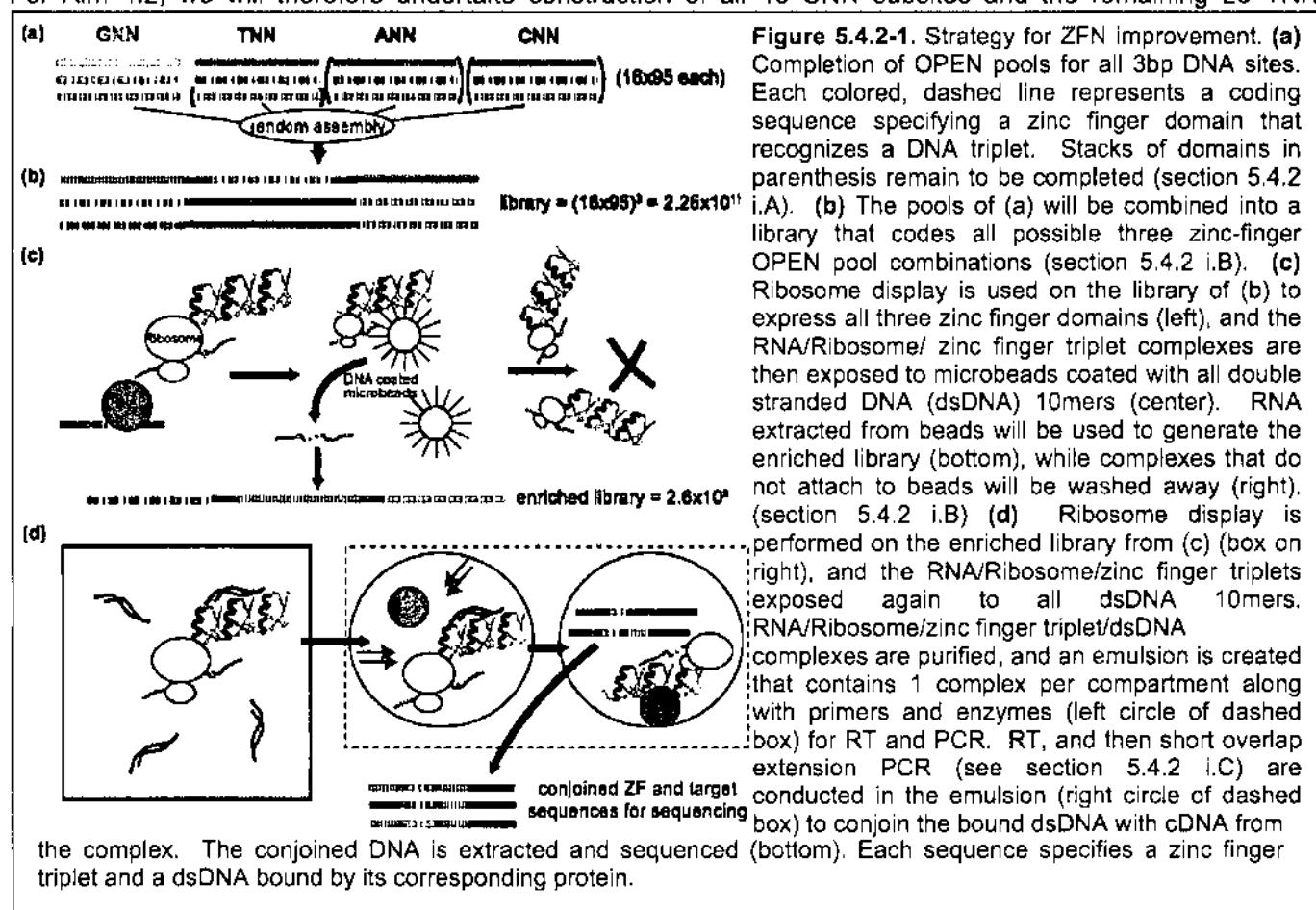


Figure 5.4.1-2: Integration of Digital Micro-mirror Device (DMD) array with the Polonator optical components. Left: Schematization of the optical path of light for DMD array control allowing selective release of cells from the Polonator flow cell. Right: Scanning Electron Microscope image of DMD mirrors and pivoting structure (Texas Instruments).

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

finger arrays for every possible 9 bp target site. Our overall strategy is described in Figure 5.4.2-1. Completion of this library will result in a very substantial advance in the ability of the academic community to engineer and use ZFNs by providing a publicly available source of pre-engineered zinc finger arrays. Currently, even in the Joung lab, where OPEN was developed, it requires 1.5 FTEs about eight weeks to perform selections for 48 ZFN half-site targets, while the cost of CompoZr™ service is high at \$25,000 per pair.

5.4.2 i.A Construction of a comprehensive set of OPEN pools: In addition to the 48 GNN and 18 TNN subsites already supported (see above), the Joung lab has already begun to construct pools for all 48 ANN subsites (16 x 3 finger positions) in collaboration with the lab of Daniel Voytas at the University of Minnesota. For Aim 4.2, we will therefore undertake construction of all 48 CNN subsites and the remaining 20 TNN



subsites. Together these additional pools will enable engineering of three-finger proteins for all possible 9 bp target sites, a substantial improvement over the current OPEN targeting range of one site every ~200 bp. Selections for the additional OPEN finger pools will proceed in a staged fashion, six at a time. Based on previous experience, we anticipate that one technician can obtain six new pools in approximately 4 weeks and therefore perform 68 selections in approximately one year. We will identify new OPEN zinc finger pools using the same randomized libraries and protocols we used to isolate the original pools (107). The existing master randomized zinc finger libraries are in a standard framework consisting of three tandem repeats of the middle finger of the murine transcription factor Zif268 in which the recognition helix residues have been altered. For each library, recognition helix residues -1, 1, 2, 3, 5, and 6 in one of the three fingers were randomized using 24 codons (degenerate sequence 5'VNS3'; V=G, A, or C; S=G or C) encoding 16 amino acids (excluding cysteine and the aromatics). The theoretical complexity of each library is therefore $24^6 = \sim 2 \times 10^8$ members. Each library has already been converted into infectious M13 phage particles as previously described (63). For each selection that yields surviving colonies, we will pick 10 clones, isolate plasmid DNA, and determine the amino acid sequences of their recognition helices. A finger pool selection will be deemed successful if the recognition helix sequences of the 10 clones resemble each other but reveal few if any identical sequences.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

We will archive successful pools of 95 clones each as previously described (107). Following successful completion of all GNN, ANN, TNN, and CNN pools, we will determine the sequences of all zinc fingers in these collections using high throughput sequencing.

Potential problems and alternative approaches: We do not anticipate significant difficulties performing selections for the 68 additional 5'TNN and 3'CNN pools because the Joung lab is the inventor of the protocol and has already successfully isolated 76 pools for 5'GNN and 5'TNN subsites. We have previously described how we handle failed selections which yield only a small number of highly similar sequences (107). If some selections for 5'TNN or 5'CNN target subsites fail to yield surviving colonies at all, we can consider: (a) Using different randomized zinc finger libraries with stronger binding "anchor" fingers (M. Maeder & J.K. Joung, unpublished observations). (b) Constructing more diverse libraries. Our current 8 year old library used 24 codons to code 16 amino acids and does not contain all possible zinc finger variants. We can now use triplet phosphoramidites (Glen Research) encoding specific codons for amino acids to make libraries with all 20 possible amino acids using only 20 codons. Not only would this library be more diverse but it would be smaller in size ($20^6=6.4 \times 10^7$ vs. $24^6 \approx 1.9 \times 10^8$). These libraries would be constructed using the Joung Lab's ABI DNA synthesizer. (c) Using different zinc finger frameworks. Instead of using the middle finger from the murine Zif268 protein as a framework for the randomized finger, we will consider alternative framework fingers from other naturally occurring proteins.

5.4.1 i.B Ribosome display to create a rarified/enriched zinc finger library (Figure 5.4.2-1 b and c): We will use ribosome display to interrogate a very large library of zinc finger variants constructed from the comprehensive set of OPEN finger pools isolated in section 5.4.2 i.A above. This large library will be constructed by randomly recombining the 64 pools for each finger position into all possible three-finger array combinations using PCR-based methods previously described by the Joung lab (63, 107). The maximum theoretical complexity of this library will be $(64 \text{ pools} \times 95 \text{ members/pool})^3 = 2.25 \times 10^{11}$, a size that could be reasonably constructed using standard ribosome display techniques (196). This large combinatorial library will be interrogated to identify members that possess specific DNA-binding activity. To do this, we will incubate the ribosome display library of zinc finger arrays with a randomized library of all possible 10 bp DNA sequences fused to magnetic beads. This randomized library will include $\sim 1.05 \times 10^5$ DNA sequences. Following equilibration for 1 hour under conditions similar to those used in phage display (69), the beads with bound zinc finger arrays will be harvested and washed with buffer to remove residual unbound proteins. RNA will be eluted from the beads, reverse transcribed into DNA, amplified by PCR, and a portion sent for high-throughput sequencing to verify that enrichment for sequences has occurred such that there are on average only 10 zinc finger arrays (vs. the original $95^3 = 8.6 \times 10^5$) derived from each of the possible $64^3=262,144$ possible combinations of finger pools. Assignment of a given zinc finger array to a particular combination of zinc finger pools will be based on the sequence information of pool clones determined in section 5.4.2 i.A above. If necessary, this selected pool of finger arrays will be converted again into a ribosome display library and interrogated with the randomized DNA site library for additional enrichment.

Potential problems and alternative approaches: We do not anticipate any difficulties constructing the large multiple pool library because the Joung lab will possess all of the required zinc finger pools and has extensive experience in building large combinatorial libraries (63, 75, 107). If the ribosome display approach described does not work, we will use the bacterial two-hybrid (B2H) selection system developed by co-I Joung as an alternative approach. This system can interrogate libraries with an upper limit of $\sim 10^9$ in size, requiring that we build and perform selections on 225 libraries composed of combinations of finger pools targeted to ~ 1200 different 9 bp sites. For each target site, we will isolate 10 zinc finger arrays from the selection. Pooling these $10 \times 262,144$ candidates will construct the equivalent of the rarified/enriched zinc finger library proposed above.

5.4.1 i.C Determining the DNA binding specificities of arrays from the rarified/enriched library (Figure 5.4.2-1d): To determine the DNA-binding specificities of the $\sim 2.6 \times 10^6$ zinc finger arrays in the rarified/enriched library, we will again perform ribosome display using the enriched library from B above. A purified solution of the RNA/ribosome/zinc finger array complexes will be created, and a mixture of all possible double stranded DNA 10mers will be added, allowing the zinc finger array complexes to bind to their DNA targets. The complexes of RNA/ribosome/zinc finger array/bound dsDNA will then be extracted. Next, we will use a variant of our emulsion PCR procedures (138) in combination with a published Short Overlap Extension (SOE) PCR protocol (57) to conjoin the RNA sequence in each complex with the dsDNA bound by the zinc finger array. In brief, an oil-water emulsion containing the purified complexes will be created such that most compartments

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

contain at most one complex. The complexes will be denatured and reverse transcriptase added to synthesize cDNA from the RNA (using procedures similar to those in section 5.3.1.1), and RNase H will be added to degrade the RNA. Primers and polymerase will then be added to enable a limited PCR reaction that, with overlaps built into the dsDNA and into the enriched library from section 5.4.1 i.B (now represented in the cDNA), will create DNA fragments in which the target dsDNA and the cDNA sequence coding for the zinc finger array are joined. The fragments will be extracted and sequenced on a next generation sequencer. Each fragment will describe a zinc finger array and a dsDNA sequence to which the array bound.

As an alternative to ribosome display, we can use an *in vitro* compartmentalization approach and next generation sequencing. *In vitro* compartmentalization will be used to couple $\sim 2.6 \times 10^7$ zinc finger arrays (to ensure 10-fold oversampling of the total sequence space) from the rarified/enriched library in section 5.4.1 i.B to 1 μm beads, using an adaptation of the published method of (51): Specifically, DNA fragments that encode fusions of finger arrays to an HA epitope tag, will be coupled to beads such that either only one or no DNA molecule is attached to each bead. These beads will also be coupled, *via* a protein A linkage, to a monoclonal antibody against the HA epitope tag. An emulsion will then be created where each droplet contains no more than a single bead, and *in vitro* transcription and translation will be performed, resulting in beads coated with both the DNA and the protein corresponding to a zinc finger array. The beads will be loaded into a Polonator machine (Preliminary Results, 4.6) flow cell and the DNA will be sequenced to identify the zinc finger arrays on each bead. The beads will then be serially interrogated with labeled clonal DNA fragments, each bearing a single 10 bp binding site, revealing the zinc finger arrays that bind the sites. While means exist to partially parallelize fragments, this system will not have the high throughput of our ribosome display technique. Another alternative is to use B2H as developed in the Joung laboratory. This is again low throughput compared to ribosome display, but we estimate that in a single selection we can fully characterize the DNA-binding specificities of 1000 zinc finger arrays, and by performing ~ 7500 such selections we can comprehensively probe the DNA-binding specificities of a very large percentage of zinc finger arrays in the rarified/enriched library.

Potential problems and alternative approaches: Variants of both ribosome display and *in vitro* compartmentalization have been tried in the context of zinc finger selections (65, 151). Among issues raised by these studies are: (a) *Non-specific binding of zinc finger arrays:* We are not overly concerned about non-specific binding because the fingers in the pools have already come through B2H selection and therefore have reasonable specificity. The enrichment step in 5.4.1 i.B can also be iterated to ensure better specificity. (b) *Non-specific binding of mRNAs to ribosomes, microbeads, dsDNAs:* We can use reverse transcriptase to double strand mRNA that is not bound to ribosomes to reduce the occurrence of RNA secondary structures that encourage non-specific mRNA binding. Studies (65, 151) each involve selection of zinc finger arrays for a single target site; thus a potential issue for our approach is (c) *Cross talk between 4^9 target sequences and 2.6×10^6 zinc finger arrays:* If cross talk proves to be an issue, we can divide the 4^9 target sequences into N pools P_1, P_2, \dots, P_N , and the zinc finger arrays into pools A_1, A_2, \dots, A_N whose predicted targets are in P_1, P_2, \dots, P_N , respectively. We can then reduce cross talk between pools P_i by using microbeads coated with double stranded DNA from the other $N-1$ pools to filter out ribosome display complexes generated from A_i . The methods of 5.4.1 i.C could then be applied within each of the N pools individually. Pools can be combined during the sequencing phase. For instance, we could create 1024 pools P_i of the form $F_1F_2F_3F_4F_5NNNN$ where each of the F_j is fixed as A, C, G, or T in a pool. To completely identify a zinc finger array and its bound target requires sequencing at most 63 bp (18bp for each of three zinc fingers + 9bp for the target) across 4-7 distinct sequence stretches.

5.4.2.ii Development of novel ZFN delivery methods: The use of ZFNs to improve engineering of human cells requires careful control of the activity of the ZFNs in the cells. Enough ZFN must be present to effect replacement of the targeted genomic element by the provided DNA template, but too much ZFN activity is cytotoxic (18). Achieving adequate control over ZFN activity can be difficult when ZFNs are generated by transfected or integrated expression constructs, as these at best enable control over timing and quantity of protein induction vs actual level or activity. An attractive alternative is to deliver calibrated quantities of externally provided ZFN proteins directly into the cells instead of trying to control their intracellular production. This can be achieved by the use of protein transduction domains (PTDs) that can penetrate directly into cells. PTDs harbor a high density of basic amino acid residues (arginine and lysine), which are critical for their transduction function (22, 76). Proteins as large as 110 kDa coupled to a PTD have been transduced into a variety of different cell types and systemic injection of such fusion proteins has demonstrated the effectiveness

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

PTD-mediated protein delivery *in vivo*. Numerous active PTDs have been described including Penetratin, polylysine, polyarginine, Tat, VP22, Syn B1, FGF-4, anthrax toxin derivative 254-amino acids (aa) peptide segment, diphtheria toxin'R' binding domain, MPG (HIV gp41/SV40 Tag NLS), pep-1, WR peptide, and exotoxin A (see references in (22, 76)).

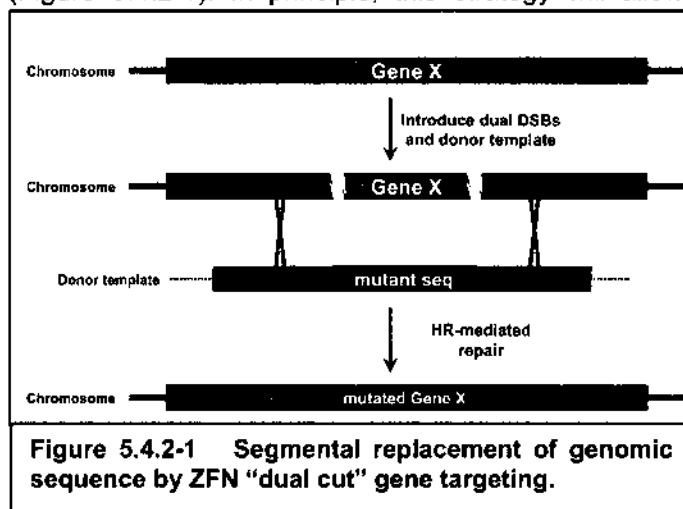
The use of PTDs to successfully improve delivery of functional Cre recombinase into mammalian cells, both *in vitro* and *in vivo*, has already been demonstrated (73, 103), including into human embryonic stem cells (127), in which recombination efficiencies of 90-100% have been reported (103, 127). We propose to adapt the procedures of (103) to determine a set of PTDs that efficiently transduce ZFNs we generate in Aims 1 and 4.1-4.2 into CTCHGV cell lines. In (103), eleven combinations of PTD domains fused to Cre were tested, and efficiency was measured by reconstitution of an inactive integrated GFP construct, after taking into account factors such as protein yield from recombinant *E. coli*, solubility, fusion protein size and charge, and the conditions and concentrations in which the fusion proteins were provided. Different PTD combinations reportedly varied by factors of as much as 8 in performance. We will apply similar procedures to a set of 5-10 ZFNs, considering as additional variables the quantity and method of delivery of template DNA provided (a component not required by Cre recombinase), using disrupted GFP reporter elements that can be repaired by the template as in section 5.1.1(iii.a).

We will also consider targeted proteolysis as a strategy for controlling intracellular ZFN levels. While PTDs control entry of ZFNs into cells, these methods control their elimination. It has recently been reported that a ZFN fused with an N-terminal degradation tag (based on ubiquitin or the FKBP12 protein) can exhibit equivalent gene targeting efficiency but less toxicity than the corresponding untagged ZFN (141). In this strategy inhibitors are used initially to suppress tag-induced degradation and open a window for ZFN operation.

Potential problems and alternatives: We do not anticipate problems testing these techniques as the methods of (103) and (141) are well documented and accessible. For targeted degradation, PROTACs (PROteolysis Targeting Chimeras) (147, 198) offer an alternative strategy. Here small molecules or peptides are used to cause a bait domain on a target protein to localize to an E3 enzyme for protein ubiquitination and degradation, an approach similar to one developed in the Church Lab whereby small molecules are used to target proteins directly to proteasome subunits (71). If consistent improvements cannot be achieved, we will rely on existing methods for ZFN expression and accept the lower efficiency that results from ZFN toxicity.

5.4.2.iii. Development of a ZFN-induced "segmental genome replacement" strategy: We propose to develop a novel sequence replacement strategy which will use two ZFN pairs to introduce a pair of DSBs flanking a region of genomic DNA to be altered. Our hypothesis is that this doubly broken stretch of genomic sequence can be repaired by a "donor template" which harbors the desired altered sequence flanked by homology arms composed of sequence adjacent to the two DSBs (Figure 5.4.2-1). We envision that if both ZFN-induced DSBs are introduced into the same allele, the cell might repair the two ends with the donor template as if they came from a single DSB, thereby replacing the original sequence between the two DSBs with the altered sequence from the donor template (Figure 5.4.2-1). In principle, this strategy will allow researchers to completely alter the sequence between the two ZFN-induced DSBs, thereby enabling more complex gene targeting alterations such as exon replacement, and to perform gene targeting even when it is not possible to design ZFNs for a target site close enough to the desired alteration site to achieve high efficiency HR.

A key requirement of this strategy is that it requires efficient introduction of plasmids encoding two ZFN pairs (four ZFN monomers) into a single cell. This can be accomplished by using vectors which express pairs of obligate heterodimeric ZFNs as a single peptide joined by a self-cleaving picornavirus T2A peptide. As noted in Preliminary Studies, section 4.5, we have built a version of such a T2A plasmid that permits rapid and easy shuttling of zinc finger arrays into this vector by simple restriction digest, and have confirmed that our vector can successfully express a functional ZFN dimer. The success of our proposed approach will depend on the efficiency with which the



Program Director/Principal Investigator (Last, First, Middle): Church, George M.

two ZFN-induced DSBs can be created on the same allele. In this regard, we note that the results cited in Preliminary Results, 4.5 show that co-expression of two pairs of ZFNs targeted to two different sites in the *HoxB13* gene led to deletion of the intervening ~180 bp sequence in ~1% to 2% of the alleles, with evidence that the deletion event is caused by NHEJ-mediated repair of the two DSB ends. This result strongly suggests that a significant percentage of alleles can be "double-cleaved" when two pairs of ZFNs are expressed in the same cell. It is likely that the efficiency of "dual cut" gene targeting will be lower than that observed with existing standard "single-cut" ZFN strategies. Here we will explore whether the long homology arms that will be generated in Aim 1 (section 5.1.1) can boost the rate of HR-mediated segment replacement.

Although the studies described in Preliminary Results, 4.5 suggest that two pairs of heterodimeric ZFNs can be expressed in a single cell, a high degree of cell death was observed in these experiments. This toxicity may be due to the formation of unwanted heterodimer species because the pair of ZFNs are not orthologous in their dimerization specificities. An important requirement for successfully developing the segmental genome replacement approach will therefore be the identification of *FokI* nuclease domains with orthologous, obligate heterodimeric interaction specificities. To create such domains, we will use a combination of iterative structure-guided design and functional testing as recently described by Miller and colleagues at Sangamo Biosciences (115).

Potential problems and alternatives: Our testing strategy is clear and unproblematic. If this segmental strategy cannot be made to work efficiently, we will rely on the improvements in ZFN HR developed in Aim 1.1. We note that the Church Lab has been developing a successful segmental genome replacement strategy in *E. coli* using selectable markers positioned near the flanks of the template DNA regions to be integrated as part of the work described in Preliminary Results, 4.4.2.

5.4.3: Aim 4.3: We will develop new high-throughput cell handling and sorting capabilities that can incorporate morphology information in addition to optical signals generated by markers, and which can operate on live cells.

In Aims 1.2 and in Aim 3 we propose to develop assays for analyzing up to millions of individual cells for genotype and allele-specific expression (ASE) information, and for *in situ* transcriptome analysis (see sections 5.1.2(ii) and 5.3.1.2). These assays require means of arraying and probing or sequencing within individual cells that are similar in nature to sequencing that is currently performed on arrayed microbeads such as in (155). The main difference is in the need to attach cells vs. microbeads to a flow cell surface and to incorporate treatments (e.g. permeabilization) to the cells that allow these targets to be accessed. Here we develop a system that supports these methods, but also extend it to incorporate aspects of the capability developed in Aim 4.1 (section 5.4.1) above, by which cells may be anchored and selectively released by light-labile chemistry and analysis-based light direction. The result will be a general purpose system for analyzing and sorting cells that enables analysis of morphology in addition to both surface and intracellular molecular content, and for selective release of cells based on morphology and content. The ability of sort cells based on morphology as well as molecular content will be significant expansion of the capabilities of FACS. Here we describe plans for: (i) arraying of cells in a flow cell for image analysis for Aims 1.2 and 3 above, which require neither selective release of cells nor that the cells be maintained alive, (ii) changes needed to support selective release, (iii) considerations needed to support live cells. Then we will describe (iv) demonstrations we will perform of these capabilities.

5.4.3 (i) arraying of cells for sections 5.1.2(ii) and 5.3.1.2: Key requirements are set by the need to use image analysis to analyze cells for sequence-related signals. This requires that the cells be present in a monolayer on a planar surface to ensure uniform focus, and that they be sufficiently well separated that image features can be assigned without error to the proper cells. While these requirements can be met minimally by immobilizing a dilute, disaggregated suspension of cells on a glass slide, our preferred approach is to array cells in a pattern on a flow cell in order to reduce incidental contacts and overlaps between cells. Techniques must be chosen carefully such that cell density and spacing are easily controlled, chemical structures (e.g. DNA and RNA) are not damaged, and attachment chemistries can withstand the forces and time (possibly 3-7 days) required for analysis. To this end, we propose to explore multiple capture and fixation techniques.

Development will begin with the construction of a flow-cell suitable for cell capture. As in Aim 4.1 (section 5.4.1), we will use our Polonator system (see Preliminary Results, 4.6) as a test instrument for development, with expansion to commercial providers of compatible instrumentation to abet technology transfer. Our current flow cell design has 8, 3.3 mm by 70 mm lanes, giving a total surface area of 1848 mm²

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

($1.848 \times 10^9 \mu\text{m}^2$). The glass surface of our flow cell will be patterned in the appropriate attachment chemistry using standard photoresist-based lithographic methods such that the attachment chemistry will appear as $5 - 10 \mu\text{m}^2$ areas with a center-to-center spacing distance of $15 \mu\text{m}$. Assuming a $10 \mu\text{m}$ cell diameter, this design will achieve a density of 1,026,666 cells per lane (> 8 million/flowcell) with $5 \mu\text{m}$ spacing between features.

We will explore multiple chemistries to anchor the cells to the flow cell, including both covalent and noncovalent attachment regimes. Each attachment chemistry has inherent advantages and disadvantages. We will test a subset of options by

arraying and fixing cells, followed by a single round of probing/sequencing and simulation of a full 4 – 7 day run, and ending with a subsequent round of probing/sequencing. Those methods that prove viable will display accurate readouts and sufficient signal intensity for both the first and last probe/sequence run with minimal loss of cells throughout the simulated run. Our preferred embodiment uses covalent attachment chemistry as they are often stronger than noncovalent interactions. The Bertozzi group recently developed a cell arraying technique that uses an azide functionalized sugar derivative displayed in polysaccharides on the cellular surface and resulted in minimal alterations to cell morphology and proliferation compared to antibody and ligand bound cells (61). The azide functionality can be used with multiple “click” chemistries to functionalize the cell surface with various attachment chemistries including ssDNA, biotin, amines, aldehydes, or direct linkage to alkynyl or phosphine derivatized surfaces (Figure 5.4.3-1, Table 5.4.3-1). (86)

Many of the proposed crosslinking chemistries have been shown to be adequate attachment chemistries for multiplexed DNA sequencing, including streptavidine/biotin, amine/amine with a homo-bifunctional NHS-ester crosslinking reagent, amine/aldehyde reductive aminations, and dsDNA formation (61, 155). While noncovalent, the hybridization of ssDNA displayed on the cell surface with that arrayed in the flowcell adds the advantage that cells can be targeted to specific portions of the array, thus allowing for multiple samples to be probed or sequenced on the same flowcell (20).

In addition to the azide-based cell capture strategies, various noncovalent attachments between arrayed antibodies or ligands have been used (e.g. concanavalin A, laminin, fibronectin) (120). These proteins can be arrayed to an aldehyde derivatized surface via reductive amination (145). However, it may be found that these molecular structures do not hold their native conformation, and thus their binding affinity, throughout the multiple heating and cooling cycles associated with high-throughput DNA sequencing (61). In addition, the high cost associated with antibodies favors click chemistry based approaches.

5.4.3 (ii) selective release of cells: To enable selective release, we functionalize the cell surface for ssDNA attachment, and use a flow cell to which complementary DNA has been attached to the surface via nitrobenzyl or other photo-labile chemistry. Release of desired cells can then be accomplished by directing 360nm light to cells, as in section 5.4.1 above.

5.4.3 (iii) considerations for use with live cells: The key requirements are: (a) Cells must be anchored to the flow cell and assayed so that the morphology or phenotype that is to be observed is not altered prior to the point of observation. (b) The conditions and duration of the assay must be sufficiently mild that the cells

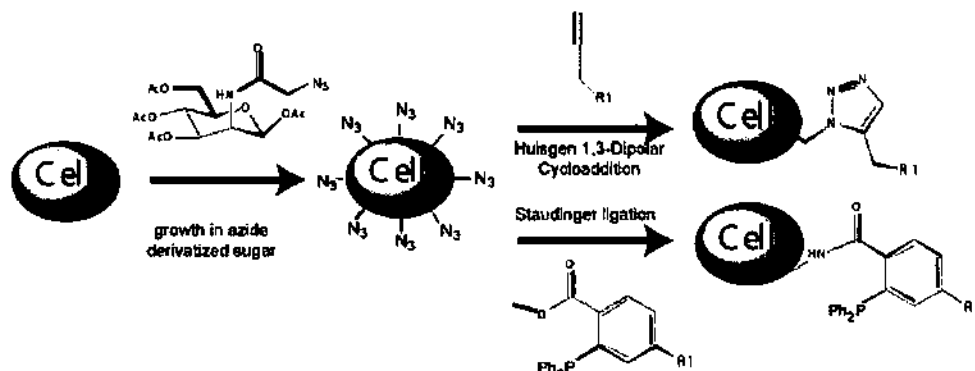


Figure 5.4.3-1. “click” chemistry capture of cells. Cells are grown in the presence of the azido sugar, which is displayed on the surface of cells. Azide groups undergo Huisgen cycloadditions or Staudinger ligations (148) to hetero-bifunctional linkers or solid surfaces. Table 5.4.3-1 describes combinations of hetero-bifunctional linkers with various functional groups (R1), surface coatings, and crosslinking agents.

Glass	R1	Reagent
streptavidine	biotin	None
Amine	Amine	BS3
Amine	aldehyde	NaBH_4
Aldehyde	Amine	NaBH_4
Alkynal	none	Huisgen
Phosphine	none	Staudinger
ssDNA	ssDNA	None

Table 5.4.3-1. Attachment chemistries

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

will survive through the point of release. These requirements exclude use of stains or hybridizations that fix or significantly perturb the cell, including many that are often used for morphological analysis of such as labeled phalloidin or anti-tubulin antibodies, which allow visualization of the actin microskelton and microtubules; however, internal structures can be visualized by use of cells containing constructs for fluorescent-protein fusions with these or other structural proteins. The ability to sort cells based on such features goes beyond FACS, which is restricted to analyzing optical properties of cells and surface-bound labels (which, of course, can also be employed by the proposed system). To constrain the times needed for analysis of potentially millions of cell images, morphological features should also be easily computable. Finally, downstream processing of live cells released and captured off the array must analyze aspects of the cell that influenced morphology at the time of observation and which will survive the subsequent process of sorting; these could be genotypes, or they could be transcriptional states that survive or regenerate after the assay. These considerations are taken into account in our proposed demonstrations:

5.4.3 (iv) proposed demonstrations: For (i) above, use of cell array and assay capabilities in support of Aims 1.2 and 3.1 (sections 5.1.2(ii) and 5.3.2.1) will comprise an actual application vs. a demonstration. For analysis and sorting of live cells (iii), we propose: (a) We will use a cell line with fluorescently labeled histone proteins and a labeled translocatable protein such as NF- κ B, and sort cells based on degree of localization of the translocatable protein. Determining the degree of translocation to the nucleus is simple from an image analysis point of view, which will shorten the duration of the assay. We will test cell sorting in two ways. (a.i) We will use cells of two different genotypes, one of which constitutively changes the level of localization of the translocatable protein, apply a mixed population to the array, sort based on localization level, and verify that sorting has successfully segregated cells by the genotypes. (a.ii) We will use cells of a single genotype and instead localize a ligand to part of the array that stimulates translocation. Here we will verify that morphology-based sorting successfully segregates cells based on the locations in which the ligand was present in the array. (b) We will attempt to recapitulate (a.i) at the level of RNA vs. genotype by using a cell population that is clonal except that a subpopulation overexpresses a factor that changes localization levels. The test will be to see if sorting successfully segregates the subset of cells that overexpress the factor, and that the RNA levels of this factor are stable through or recover from the conditions of the assay.

Potential problems and alternatives: A key issue for live cells is that they must be allowed to attach via their native mechanisms to suitable ligands to avoid anoikis. This is accomplished above by use of native ligands attached to the surface via a photocleavable substrate (preferably DNA, but also alternatives given in 5.4.2 (ii)). However, once live cells are on the surface, they may begin to migrate and attach to each other. To avoid this, we will lay down the ligand in grids. If this is insufficient, we will explore ways of generating direct cross-links between cell surface proteins and the surface, in effect "leashing" the cells to their locations.

Aim 4 goals: Final goals Our targets are: For Aim 4.1, we will synthesize 1000 complete ZFN proteins using the platform we develop. For Aim 4.2, we will complete the OPEN zinc finger pools and characterize the binding specificity of 10,000 triplet zinc finger arrays from the rarified library using the described ribosome display system. For Aim 4.3, we will build the image analysis and cell arraying required for Aims 1.2 and 3 on the Polonator, and demonstrate cell sorting by morphology. **Intermediate goals** As noted in our Research Design Overview, we will evaluate progress at the end of year 2 of the Center and renegotiate goals as appropriate. By that time we expect that: *Aim 4.1:* We will have demonstrated the ability to selectively release and capture microbeads based on DNA sequences on them required of the system in Aim 4.1. *Aim 4.2:* We expect to have completed all of the OPEN pool selections, and to have successfully tested the compartmentalized SOE-PCR that conjoins RNA bound to ribosomes with target DNA of zinc finger triplets, for simple mixtures of triplets vs the full combinatorial library. *Aim 4.3:* Cell handling for Aim 1.3 and 3 will have been completed. **Impacts:** The integrated DNA sequencing and synthesis platform will greatly improve the ability to create complex libraries of large DNA constructs and will likely be adopted commercially. Completion of the OPEN zinc finger pools and development of improved ZFN targeting and delivery techniques will put effective human cell genetic engineering in the hands of the research community, where it will broadly support biomedical research generally, and gene therapy in particular. The ribosome display experiments of Aim 4.2 will generate an extremely large data base of zinc finger array specificities and will provide unparalleled opportunities to develop new computational methods for designing arrays with new binding specificities. Extending cell sorting technology to incorporate cell morphology along with cell staining characteristics will be the basis of a new form of high-throughput screening that will have broad application in research.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Bibliography and References

MGI CEGS Publication Bibliography

The following is a list of publications of the MGI Center broken down by category.

- MGIC 2004-2009 publications (excludes MGI first year (2003) publications): 44
- MGIC submitted: 1
- MGIC conference proceedings: 1
- MGIC electronic : 1
- MGIC 2003 publications: 8

MGIC 2004-2009 publications (44)

- Aach, J. and G. M. Church (2004). "Mathematical models of diffusion-constrained polymerase chain reactions: basis of high-throughput nucleic acid assays and simple self-organizing systems." *J Theor Biol* **228**(1): 31-46.
- Bakal, C., J. Aach, G. Church and N. Perrimon (2007). "Quantitative morphological signatures define local signaling networks regulating cell morphology." *Science* **316**(5832): 1753-6.
- Ball, M. P., J. B. Li, Y. Gao, J. Lee, E. LeProust, I.-H. Park, B. Xie, G. Q. Daley and G. M. Church (2009). "Targeted and whole-genome methylomics reveals gene-body signatures in human cell lines." *Nat Biotechnol* **27**(4): 361-368.
- Church, G. M. (2006). "Genomes for all." *Sci Am* **294**(1): 46-54.
- Church, G. M., G. J. Porreca, R. C. Terry and M. Lares (2008). "High-Speed Imaging for DNA Sequencing." *Biophotonics* (<http://www.photonics.com/Content/ReadArticle.aspx?ArticleID=33989>).
- Conrad, C., J. Zhu, C. Conrad, D. Schoenfeld, Z. Fang, M. Ingelsson, S. Stamm, G. Church and B. T. Hyman (2007). "Single molecule profiling of tau gene expression in Alzheimer's disease." *J Neurochem* **103**(3): 1228-36.
- Dantas, G., M. O. Sommer, R. D. Oluwasegun and G. M. Church (2008). "Bacteria subsisting on antibiotics." *Science* **320**(5872): 100-3.
- Kim, D. S., S. E. Ross, J. M. Trimarchi, J. Aach, M. E. Greenberg and C. L. Cepko (2008). "Identification of molecular markers of bipolar cells in the murine retina." *J Comp Neurol* **507**(5): 1795-810.
- Kim, J. B., G. J. Porreca, L. Song, S. C. Greenway, J. M. Gorham, G. M. Church, C. E. Seidman and J. G. Seidman (2007). "Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy." *Science* **316**(5830): 1481-4.
- Lee, C. S., L. Y. Tee, S. Dusenbery, T. Takata, J. P. Golden, B. A. Pierchala, D. I. Gottlieb, E. M. Johnson, Jr., D. W. Choi and B. J. Snider (2005). "Neurotrophin and GDNF family ligands promote survival and alter excitotoxic vulnerability of neurons derived from murine embryonic stem cells." *Exp Neurol* **191**(1): 65-76.
- Lee, H. S., J. L. Sherley, J. J. Chen, C. C. Chiu, L. L. Chiou, J. D. Liang, P. C. Yang, G. T. Huang and J. C. Sheu (2005). "EMP-1 is a junctional protein in a liver stem cell line and in the liver." *Biochem Biophys Res Commun* **334**(4): 996-1003.
- Lee, S. I., D. Pe'er, A. M. Dudley, G. M. Church and D. Koller (2006). "Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification." *Proc Natl Acad Sci U S A* **103**(38): 14062-7. PMC ID: PMC1599912.
- Leparc, G. G. and R. D. Mitra (2007). "Non-EST-based prediction of novel alternatively spliced cassette exons with cell signaling function in *Caenorhabditis elegans* and human." *Nucleic Acids Res* **35**(10): 3192-202. PMC ID: PMC1904267.
- Leparc, G. G. and R. D. Mitra (2007). "A sensitive procedure to detect alternatively spliced mRNA in pooled-tissue samples." *Nucleic Acids Res* **35**(21): e146. PMC ID: PMC2175357.
- Li, J. B., Y. Gao, J. Aach, K. Zhang, G. V. Kryukov, B. Xie, A. Ahlford, J.-K. Yoon, A. M. Rosenbaum, A. Wait-Zaranek, E. LeProust, S. Sunyaev and G. M. Church (2009). "Multiplex padlock capture and sequencing reveal human hypermutable CpG variations." *Genome Res*: in press.
- Li, J. B., E. Y. Levanon, J.-K. Yoon, J. Aach, B. Xie, E. LeProust, K. Zhang, Y. Gao and G. M. Church (2009). "Genome-wide Identification of Human RNA Editing Sites by Parallel DNA Capturing and Sequencing." *Science*: in press.
- Lunshof, J. E. (2006). "Personalized medicine: new perspectives – new ethics?" *Personalized Med* **3**(2): 187-194.
- Mikkilineni, V., R. D. Mitra, J. Merritt, J. R. DiTonno, G. M. Church, B. Ogunnaike and J. S. Edwards (2004). "Digital quantitative measurements of gene expression." *Biotechnol Bioeng* **86**(2): 117-24.
- Nardi, V., T. Raz, X. Cao, C. J. Wu, R. M. Stone, J. Cortes, M. W. Deininger, G. Church, J. Zhu and G. Q. Daley (2008). "Quantitative monitoring by polymerase colony assay of known mutations resistant to ABL kinase inhibitors." *Oncogene* **27**(6): 775-82.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

- Pare, J. F. and J. L. Sherley (2006). "Biological principles for ex vivo adult stem cell expansion." *Curr Top Dev Biol* **73**: 141-71.
- Porreca, G. J., J. Shendure and G. M. Church (2006). "Polony DNA sequencing." *Curr Protoc Mol Biol* **Chapter 7**: Unit 7 8.
- Porreca, G. J., K. Zhang, J. B. Li, B. Xie, D. Austin, S. L. Vassallo, E. M. LeProust, B. J. Peck, C. J. Emig, F. Dahl, Y. Gao, G. M. Church and J. Shendure (2007). "Multiplex amplification of large sets of human exons." *Nat Methods* **4**(11): 931-6.
- Rambhalla, L., S. Ram-Mohan, J. J. Cheng and J. L. Sherley (2005). "Immortal DNA strand cosegregation requires p53/IMPDH-dependent asymmetric self-renewal associated with adult stem cells." *Cancer Res* **65**(8): 3155-61.
- Rieger, C., R. Poppino, R. Sheridan, K. Moley, R. Mitra and D. Gottlieb (2007). "Polony analysis of gene expression in ES cells and blastocysts." *Nucleic Acids Res* **35**(22): e151. PMC ID: PMC2190707.
- Schwartz, D., M. F. Chou and G. M. Church (2009). "Predicting protein post-translational modifications using meta-analysis of proteome scale data sets." *Mol Cell Proteomics* **8**(2): 365-79. PMC ID: PMC2634583.
- Shendure, J., R. D. Mitra, C. Varma and G. M. Church (2004). "Advanced sequencing technologies: methods and goals." *Nat Rev Genet* **5**(5): 335-44.
- Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra and G. M. Church (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." *Science* **309**(5741): 1728-32.
- Shendure, J. A., G. J. Porreca and G. M. Church (2008). "Overview of DNA sequencing strategies." *Curr Protoc Mol Biol* **Chapter 7**: Unit 7 1.
- Sherley, J. L. (2007). "Commentary: Facing up to the feasibility of ANT-OAR." *Stem Cell Rev* **3**(1): 66-7.
- Sherley, J. L. (2008). "All good cells come from cells." *Nat Cell Biol* **10**(3): 248.
- Tannenbaum, E., J. L. Sherley and E. I. Shakhnovich (2004). "Imperfect DNA lesion repair in the semiconservative quasispecies model: derivation of the Hamming class equations and solution of the single-fitness peak landscape." *Phys Rev E Stat Nonlin Soft Matter Phys* **70**(6 Pt 1): 061915.
- Tannenbaum, E., J. L. Sherley and E. I. Shakhnovich (2005). "Evolutionary dynamics of adult stem cells: comparison of random and immortal-strand segregation mechanisms." *Phys Rev E Stat Nonlin Soft Matter Phys* **71**(4 Pt 1): 041914.
- Tannenbaum, E., J. L. Sherley and E. I. Shakhnovich (2006). "Semiconservative quasispecies equations for polysomic genomes: the haploid case." *J Theor Biol* **241**(4): 791-805.
- Turner, D. J., J. Shendure, G. Porreca, G. Church, P. Green, C. Tyler-Smith and M. E. Hurles (2006). "Assaying chromosomal inversions by single-molecule haplotyping." *Nat Methods* **3**(6): 439-45.
- Vigneault, F., A. M. Sismour and G. M. Church (2008). "Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation." *Nat Methods* **5**(9): 777-779.
- Wang, H., M. Johnston and R. D. Mitra (2007). "Calling cards for DNA-binding proteins." *Genome Res* **17**(8): 1202-9. PMC ID: PMC1933518.
- Wei, L., L. Cui, B. J. Snider, M. Rivkin, S. S. Yu, C. S. Lee, L. D. Adams, D. I. Gottlieb, E. M. Johnson, Jr., S. P. Yu and D. W. Choi (2005). "Transplantation of embryonic stem cells overexpressing Bcl-2 promotes functional recovery after transient cerebral ischemia." *Neurobiol Dis* **19**(1-2): 183-93.
- Willerth, S. M., K. J. Arendas, D. I. Gottlieb and S. E. Sakiyama-Elbert (2006). "Optimization of fibrin scaffolds for differentiation of murine embryonic stem cells into neural lineage cells." *Biomaterials* **27**(36): 5990-6003. PMC ID: PMC1794024.
- Willerth, S. M., T. E. Fixel, D. I. Gottlieb and S. E. Sakiyama-Elbert (2007). "The effects of soluble growth factors on embryonic stem cell differentiation inside of fibrin scaffolds." *Stem Cells* **25**(9): 2235-44. PMC ID: PMC2637150.
- Xian, H. and D. I. Gottlieb (2004). "Dividing Olig2-expressing progenitor cells derived from ES cells." *Glia* **47**(1): 88-101.
- Xian, H. Q., K. Werth and D. I. Gottlieb (2005). "Promoter analysis in ES cell-derived neural cells." *Biochem Biophys Res Commun* **327**(1): 155-62.
- Zaraneek, A. W., W. Clegg, Vandewege and G. M. Church (2008). *Free Factories: Unified Infrastructure for Data Intensive Web Services* USENIX Annual Technical Conference, Boston, MA.
- Zhang, K., A. C. Martiny, N. B. Reppas, K. W. Barry, J. Malek, S. W. Chisholm and G. M. Church (2006). "Sequencing genomes from single cells by polymerase cloning." *Nat Biotechnol* **24**(6): 680-6.
- Zhang, K., J. Zhu, J. Shendure, G. J. Porreca, J. D. Aach, R. D. Mitra and G. M. Church (2006). "Long-range polony haplotyping of individual human chromosome molecules." *Nat Genet* **38**(3): 382-7.

MGIC submitted (1)

Unpublished

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

MGIC conference proceedings (1)

Forest, C. R., S. E. Ross and G. M. Church (2008). DNA Sequencing By Ligation On Surface-Bound Beads In A Microchannel Environment (Paper ID No. 0261). microTAS, San Diego.

MGIC electronic(1)

Terry, R., G. Porreca, K. McCarthy and G. M. Church (2008) "Polonator Instrument <http://www.polonator.org> "

MGIC 2003 publications (8)

- Adams, L. D., L. Choi, H. Q. Xian, A. Yang, B. Sauer, L. Wei and D. I. Gottlieb (2003). "Double lox targeting for neural cell transgenesis." *Brain Res Mol Brain Res* **110**(2): 220-33.
- Lee, H. S., G. G. Crane, J. R. Merok, J. R. Tunstead, N. L. Hatch, K. Panchalingam, M. J. Powers, L. G. Griffith and J. L. Sherley (2003). "Clonal expansion of adult rat hepatic stem cell lines by suppression of asymmetric cell kinetics (SACK)." *Biotechnol Bioeng* **83**(7): 760-71.
- Merritt, J., J. R. DiTonno, R. D. Mitra, G. M. Church and J. S. Edwards (2003). "Parallel competition analysis of *Saccharomyces cerevisiae* strains differing by a single base using polymerase colonies." *Nucleic Acids Res* **31**(15): e84. PMC ID: PMC169973.
- Mitra, R. D., V. L. Butty, J. Shendure, B. R. Williams, D. E. Housman and G. M. Church (2003). "Digital genotyping and haplotyping with polymerase colonies." *Proc Natl Acad Sci U S A* **100**(10): 5926-31. PMC ID: PMC156303.
- Mitra, R. D., J. Shendure, J. Olejnik, O. Edyta Krzymanska and G. M. Church (2003). "Fluorescent in situ sequencing on polymerase colonies." *Anal Biochem* **320**(1): 55-65.
- Qu, Y., S. Vadivelu, L. Choi, S. Liu, A. Lu, B. Lewis, R. Girgis, C. S. Lee, B. J. Snider, D. I. Gottlieb and J. W. McDonald (2003). "Neurons derived from embryonic stem (ES) cells resemble normal neurons in their vulnerability to excitotoxic death." *Exp Neurol* **184**(1): 326-36.
- Xian, H. Q., E. McNichols, A. St Clair and D. I. Gottlieb (2003). "A subset of ES-cell-derived neural cells marked by gene targeting." *Stem Cells* **21**(1): 41-9.
- Zhu, J., J. Shendure, R. D. Mitra and G. M. Church (2003). "Single molecule profiling of alternative pre-mRNA splicing." *Science* **301**(5634): 836-8.

References cited in this proposal

1. Aasen T, Raya A, Barrero MJ, Garreta E, Consiglio A, Gonzalez F, Vassena R, Bilic J, Pekarik V, Tiscornia G, Edel M, Boue S, Belmonte JC. 2008. Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat Biotechnol* 26: 1276-84.
2. Aiuti A, Bachoud-Levi AC, Blesch A, Brenner MK, Cattaneo F, Chioccia EA, Gao G, High KA, Leen AM, Lemoine NR, McNeish IA, Meneguzzi G, Peschanski M, Roncarolo MG, Strayer DS, Tuszynski MH, Waxman DJ, Wilson JM. 2007. Progress and prospects: gene therapy clinical trials (part 2). *Gene Ther* 14: 1555-63.
3. Alexander BL, Ali RR, Alton EW, Bainbridge JW, Braun S, Cheng SH, Flotte TR, Gaspar HB, Grez M, Griesenbach U, Kaplitt MG, Ott MG, Seger R, Simons M, Thrasher AJ, Thrasher AZ, Yla-Herttuala S. 2007. Progress and prospects: gene therapy clinical trials (part 1). *Gene Ther* 14: 1439-47.
4. Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* 322: 881-8.
5. Anderson P, Kedersha N. 2006. RNA granules. *J Cell Biol* 172: 803-8. PMC ID: PMC2063724.
6. Anderson P, Kedersha N. 2008. Stress granules: the Tao of RNA triage. *Trends Biochem Sci* 33: 141-50.
7. Bakal C, Aach J, Church G, Perrimon N. 2007. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science* 316: 1753-6.
8. Unpublished
9. Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7: 781-91.
10. Ball MP, Li JB, Gao Y, Lee J, LeProust E, Park I-H, Xie B, Daley GQ, Church GM. 2009. Targeted and whole-genome methylomics reveals gene-body signatures in human cell lines. *Nat Biotechnol* 27: 361-8.
11. Bang D, Church GM. 2008. Gene synthesis by circular assembly amplification. *Nat Methods* 5: 37-9.
12. Behlke MA, Devor EJ. 2005. *Chemical Synthesis of Oligonucleotides* (http://www.idtdna.com/Support/Technical/TechnicalBulletinPDF/Chemical_Synthesis_of_Oligonucleotides.pdf), Integrated DNA Technologies

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

13. Bell J. 2004. Predicting disease using genomics. *Nature* 429: 453-6.
14. Bitinaite J, Wah DA, Aggarwal AK, Schildkraut I. 1998. FokI dimerization is required for DNA cleavage. *Proc Natl Acad Sci U S A* 95: 10570-5. PMC ID: PMC27935.
15. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18: 630-4.
16. Brenner S, Williams SR, Vermaas EH, Storck T, Moon K, McCollum C, Mao J, Luo S, Kirchner JJ, Eletr S, DuBridge RB, Burcham T, Albrecht G. 2000. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc Natl Acad Sci U S A* 97: 1665-70. PMC ID: PMC26493.
17. Carr PA, Park JS, Lee YJ, Yu T, Zhang S, Jacobson JM. 2004. Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res* 32: e162. PMC ID: PMC534640.
18. Cathomen T, Joung JK. 2008. Zinc-finger nucleases: the next generation emerges. *Mol Ther* 16: 1200-7.
19. Cavazzana-Calvo M, Fischer A. 2007. Gene therapy for severe combined immunodeficiency: are we there yet? *J Clin Invest* 117: 1456-65. PMC ID: PMC1878528.
20. Chandra RA, Douglas ES, Mathies RA, Bertozzi CR, Francis MB. 2006. Programmable Cell Adhesion Encoded by DNA Hybridization *Angew. Chem. Int.* 45: 896-901.
21. Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 56: 18-31.
22. Chauhan A, Tikoo A, Kapur AK, Singh M. 2007. The taming of the cell penetrating domain of the HIV Tat: myths and realities. *J Control Release* 117: 148-62. PMC ID: PMC1859861.
23. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE. 2008. Variations in DNA elucidate molecular networks that cause disease. *Nature* 452: 429-35.
24. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437: 1365-9.
25. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 6: 99-103. PMC ID: PMC2630795.
26. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, De Jager PL, Shaw SY, Wolfish CS, Slavik JM, Cotsapas C, Rivas M, Dermizakis ET, Cahir-McFarland E, Kieff E, Hafler D, Daly MJ, Altshuler D. 2008. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet* 4: e1000287. PMC ID: PMC2583954.
27. Christian AT, Pattee MS, Attix CM, Reed BE, Sorensen KJ, Tucker JD. 2001. Detection of DNA point mutations and mRNA expression levels by rolling circle amplification in individual cells. *Proc Natl Acad Sci U S A* 98: 14238-43. PMC ID: PMC64666.
28. Church GM. 2006. Genomes for all. *Sci Am* 294: 46-54.
29. Church GM, Porreca GJ, Terry RC, Lares M. 2008. High-Speed Imaging for DNA Sequencing. *Biophotonics* (<http://www.photonics.com/Content/ReadArticle.aspx?ArticleID=33989>).
30. Claassen DA, Desler MM, Rizzino A. 2009. ROCK inhibition enhances the recovery and growth of cryopreserved human embryonic stem cells and human induced pluripotent stem cells. *Mol Reprod Dev*.
31. Conrad C, Gupta R, Mohan H, Niess H, Bruns CJ, Kopp R, von Luetichau I, Guba M, Heeschen C, Jauch KW, Huss R, Nelson PJ. 2007. Genetically engineered stem cells for therapeutic gene delivery. *Curr Gene Ther* 7: 249-60.
32. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10: 184-94.
33. Costantino N, Court DL. 2003. Enhanced levels of lambda Red-mediated recombinants in mismatch repair mutants. *Proc Natl Acad Sci U S A* 100: 15748-53. PMC ID: PMC307639.
34. Datta S, Costantino N, Zhou X, Court DL. 2008. Identification and analysis of recombineering functions from Gram-negative and Gram-positive bacteria and their phages. *Proc Natl Acad Sci U S A* 105: 1626-31. PMC ID: PMC2234195.
35. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat Genet* 37: 1217-23.
36. Dekker M, Brouwers C, te Riele H. 2003. Targeted gene modification in mismatch-repair-deficient embryonic stem cells by single-stranded DNA oligonucleotides. *Nucleic Acids Res* 31: e27. PMC ID: PMC152881.
37. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y, Zhang K. 2009. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* 27: 353-60.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

38. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Luan L, Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Gudnucchi C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chim GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316: 1331-6.
39. Ding Y. 2006. Statistical and Bayesian approaches to RNA secondary structure prediction. *RNA* 12: 323-31. PMC ID: PMC1383571.
40. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO. 2007. A genome-wide association study of global gene expression. *Nat Genet* 39: 1202-7.
41. Donoho G, Jasin M, Berg P. 1998. Analysis of gene targeting and intrachromosomal homologous recombination stimulated by genomic double-strand breaks in mouse embryonic stem cells. *Mol Cell Biol* 18: 4070-8. PMC ID: PMC108991.
42. Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, Hocking TD, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Amacher SL. 2008. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat Biotechnol* 26: 702-8. PMC ID: PMC2674762.
43. Drubin DA, Way JC, Silver PA. 2007. Designing biological systems. *Genes Dev* 21: 242-54.
44. Eberwine J, Kacharmina JE, Andrews C, Miyashiro K, McIntosh T, Becker K, Barrett T, Hinkle D, Dent G, Marciano P. 2001. mRNA expression analysis of tissue sections and single cells. *J Neurosci* 21: 8310-4.
45. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Veceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133-8.
46. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K. 2008. Genetics of gene expression and its effect on disease. *Nature* 452: 423-8.
47. Eulalia A, Behm-Ansmant I, Izaurralde E. 2007. P bodies: at the crossroads of post-transcriptional pathways. *Nat Rev Mol Cell Biol* 8: 9-22.
48. Foley JE, Yeh JR, Maeder ML, Reyon D, Sander JD, Peterson RT, Joung JK. 2009. Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN). *PLoS ONE* 4: e4348. PMC ID: PMC2634973.
49. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316: 889-94. PMC ID: PMC2646098.
50. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science* 318: 1136-40.
51. Griffiths AD, Tawfik DS. 2006. Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol* 24: 395-402.
52. Grillot-Courvalin C, Goussard S, Huetz F, Ojcius DM, Courvalin P. 1998. Functional gene transfer from intracellular bacteria to mammalian cells. *Nat Biotechnol* 16: 862-6.
53. Guccione E, Martinato F, Finocchiaro G, Luzi L, Tizzoni L, Dall' Olio V, Zardo G, Nervi C, Bernard L, Amati B. 2006. Myc-binding-site recognition in the human genome is determined by chromatin context. *Nat Cell Biol* 8: 764-70.
54. Hanna J, Wernig M, Markoulaki S, Sun CW, Meissner A, Cassady JP, Beard C, Brambrink T, Wu LC, Townes TM, Jaenisch R. 2007. Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science* 318: 1920-3.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

55. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z. 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320: 106-9.
56. Hayden EC. 2008. Give me my genome (October 21, 2008). In *Nature* (<http://www.nature.com.ezp-prod1.hul.harvard.edu/news/2008/081021/full/news.2008.1182.html>)
57. Heckman KL, Pease LR. 2007. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat Protoc* 2: 924-32.
58. Heinemann JA, Sprague GF, Jr. 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* 340: 205-9.
59. Hindorf LA, Junkins HA, Mehta JP, Manolio TA. 2009. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/26525384. Accessed May 12, 2009.
60. Holden C. 2008. Genomes for the world. *Science* 322: 509.
61. Hsiao SC, Crow AK, Lam WA, Bertozzi CR, Fletcher DA, Francis MB. 2008. DNA-Coated AFM Cantilevers for the Investigation of Cell Adhesion and the Patterning of Live Cells. *Angew. Chem. Int.* 47: 8473-7.
62. Huangfu D, Osafune K, Maeht R, Guo W, Eijkelenboom A, Chen S, Muhlestein W, Melton DA. 2008. Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nat Biotechnol* 26: 1269-75.
63. Hurt JA, Thibodeau SA, Hirsh AS, Pabo CO, Joung JK. 2003. Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection. *Proc Natl Acad Sci U S A* 100: 12271-6. PMC ID: PMC218748.
64. Igoucheva O, Alexeev V, Yoon K. 2004. Oligonucleotide-directed mutagenesis and targeted gene correction: a mechanistic point of view. *Curr Mol Med* 4: 445-63.
65. Ihara H, Mie M, Funabashi H, Takahashi F, Sawasaki T, Endo Y, Kobatake E. 2006. In vitro selection of zinc finger DNA-binding proteins through ribosome display. *Biochem Biophys Res Commun* 345: 1149-54.
66. Illumina Corp. 2009. Illumina Presents Development Roadmap for Scaling its Genome Analyzer Innovations to substantially increase output, decrease cost, and expand applications (<http://investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&ID=1252407&highlight=>).
67. International 1000 Genomes Consortium. 1000 Genomes Project (<http://www.1000genomes.org/>).
68. International HapMap C. 2005. A haplotype map of the human genome. *Nature* 437: 1299-320. PMC ID: PMC1880871.
69. Isalan M, Choo Y. 2001. Engineering nucleic acid-binding proteins by phage display. *Methods Mol Biol* 148: 417-29.
70. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J, Jr., Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283: 83-7.
71. Janse DM, Crosas B, Finley D, Church GM. 2004. Localization to the proteasome is sufficient for degradation. *J Biol Chem* 279: 21415-20.
72. Jantz D, Amann BT, Gatto GJ, Jr., Berg JM. 2004. The design of functional DNA-binding proteins based on zinc finger domains. *Chem Rev* 104: 789-99.
73. Jo D, Nashabi A, Doxsee C, Lin Q, Unutmaz D, Chen J, Ruley HE. 2001. Epigenetic regulation of gene structure and function with a cell-permeable Cre recombinase. *Nat Biotechnol* 19: 929-33.
74. Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316: 1497-502.
75. Joung JK, Ramm EI, Pabo CO. 2000. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc Natl Acad Sci U S A* 97: 7382-7. PMC ID: PMC16554.
76. Kabouridis PS. 2003. Biological applications of protein transduction technology. *Trends Biotechnol* 21: 498-503. PMC ID: PMC2597147.
77. Kaji K, Norrby K, Paca A, Mileikovsky M, Mohseni P, Woltjen K. 2009. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature*.
78. Kang HM, Ye C, Eskin E. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180: 1909-25. PMC ID: PMC2600931.
79. Kelly JR, Rubin AJ, Davis JH, Ajo-Franklin CM, Cumbers J, Czar MJ, de Mora K, Glieberman AL, Monie DD, Endy D. 2009. Measuring the Activity of BioBrick Promoters Using an In Vivo Reference Standard. *J Biol Eng* 3: 4.
80. Kendzierski CM, Chen M, Yuan M, Lan H, Attie AD. 2006. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* 62: 19-27.
81. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* 12: 996-1006. PMC ID: PMC186604.
82. Kharchenko PV, Tolstoukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 26: 1351-9. PMC ID: PMC2597701.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

83. Kim DS, Ross SE, Trimarchi JM, Aach J, Greenberg ME, Cepko CL. 2008. Identification of molecular markers of bipolar cells in the murine retina. *J Comp Neurol* 507: 1795-810.
84. Kim JB, Porreca GJ, Song L, Greenway SC, Gorham JM, Church GM, Seidman CE, Seidman JG. 2007. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* 316: 1481-4.
85. Klein RJ. 2007. Power analysis for genome-wide association studies. *BMC Genet* 8: 58. PMC ID: PMC2042984.
86. Kolb HC, Finn MG, B. SK. 2001. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem. Int.* 40: 2004-21.
87. Kouprina N, Campbell M, Graves J, Campbell E, Meincke L, Tesmer J, Grady DL, Doggett NA, Moyzis RK, Deaven LL, Larionov V. 1998. Construction of human chromosome 16- and 5-specific circular YAC/BAC libraries by in vivo recombination in yeast (TAR cloning). *Genomics* 53: 21-8.
88. Kouprina N, Kawamoto K, Barrett JC, Larionov V, Koi M. 1998. Rescue of targeted regions of mammalian chromosomes by in vivo recombination in yeast. *Genome Res* 8: 666-72. PMC ID: PMC310736.
89. Kouprina N, Larionov V. 2006. Selective isolation of mammalian genes by TAR cloning. *Curr Protoc Hum Genet* Chapter 5: Unit 5 17.
90. Kouprina N, Larionov V. 2008. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast *Saccharomyces cerevisiae*. *Nat Protoc* 3: 371-7.
91. Kouprina N, Leem SH, Solomon G, Ly A, Koriabine M, Otstot J, Pak E, Dutra A, Zhao S, Barrett JC, Larionov V. 2003. Segments missing from the draft human genome sequence can be isolated by transformation-associated recombination cloning in yeast. *EMBO Rep* 4: 257-62. PMC ID: PMC1315894.
92. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324: 255-8.
93. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* 37: D755-61.
94. Kurimoto K, Yabuta Y, Ohinata Y, Saitou M. 2007. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat Protoc* 2: 739-52.
95. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* 40: 225-31.
96. Kwiatkowski M, Fredriksson S, Isaksson A, Nilsson M, Landegren U. 1999. Inversion of in situ synthesized oligonucleotides: improved reagents for hybridization and primer extension in DNA microarrays. *Nucleic Acids Res* 27: 4710-4. PMC ID: PMC148770.
97. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

- JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing C. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
98. Laner A, Goussard S, Ramalho AS, Schwarz T, Amaral MD, Courvalin P, Schindelhauer D, Grillot-Courvalin C. 2005. Bacterial transfer of large functional genomic DNA into human cells. *Gene Ther* 12: 1559-72.
99. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18: 1851-8. PMC ID: PMC2577856.
100. Li JB, Gao Y, Aach J, Zhang K, Kryukov GV, Xie B, Ahlford A, Yoon J-K, Rosenbaum AM, Wait-Zaranek A, LeProust E, Sunyaev S, Church GM. 2009. Multiplex padlock capture and sequencing reveal human hypermutable CpG variations. *Genome Res* in press.
101. Li JB, Levanon EY, Yoon J-K, Aach J, Xie B, LeProust E, Zhang K, Gao Y, G.M. C. 2009. Genome-wide Identification of Human RNA Editing Sites by Parallel DNA Capturing and Sequencing. *Science* in press.
102. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazendonk E, Prins P, Plasterk RH, Jansen RC, Breitling R, Kammenga JE. 2006. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet* 2: e222. PMC ID: PMC1756913.
103. Lin Q, Jo D, Gebre-Amlak KD, Ruley HE. 2004. Enhanced cell-permeant Cre protein for site-specific recombination in cultured cells. *BMC Biotechnol* 4: 25. PMC ID: PMC529453.
104. Link AJ, Phillips D, Church GM. 1997. Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization. *J Bacteriol* 179: 6228-37. PMC ID: PMC179534.
105. Lufino MM, Edser PA, Wade-Martins R. 2008. Advances in high-capacity extrachromosomal vector technology: episomal maintenance, vector delivery, and transgene expression. *Mol Ther* 16: 1525-38.
106. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. 2008. From genetic privacy to open consent. *Nat Rev Genet* 9: 406-11.
107. Maeder ML, Thibodeau-Beganny S, Osiaik A, Wright DA, Anthony RM, Eichinger M, Jiang T, Foley JE, Winfrey RJ, Townsend JA, Unger-Wallace E, Sander JD, Muller-Lerch F, Fu F, Pearlberg J, Gobel C, Dassié JP, Pruett-Miller SM, Porteus MH, Sgroi DC, Iafrate AJ, Dobbs D, McCray PB, Jr., Cathomen T, Voytas DF, Joung JK. 2008. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* 31: 294-301. PMC ID: PMC2535758.
108. Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, Stadtfeld M, Yachechko R, Tchieu J, Jaenisch R, Plath K, Hochedlinger K. 2007. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* 1: 55-70.
109. Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288: 911-40.
110. Maynard ND, Chen J, Stuart RK, Fan JB, Ren B. 2008. Genome-wide mapping of allele-specific protein-DNA interactions in human cells. *Nat Methods* 5: 307-9.
111. McCarroll SA. 2008. Extending genome-wide association studies to copy-number variation. *Hum Mol Genet* 17: R135-42.
112. Meng Q, Kim DH, Bai X, Bi L, Turro NJ, Ju J. 2006. Design and synthesis of a photocleavable fluorescent nucleotide 3'-O-allyl-dGTP-PC-Bodipy-FL-510 as a reversible terminator for DNA sequencing by synthesis. *J Org Chem* 71: 3248-52.
113. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448: 553-60.
114. Milani L, Lundmark A, Nordlund J, Kitalainen A, Flaegstad T, Jonmundsson G, Kanerva J, Schmiegelow K, Gunderson KL, Lonnnerholm G, Syvanen AC. 2009. Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome Res* 19: 1-11. PMC ID: PMC2612957.
115. Miller JC, Holmes MC, Wang J, Guschin DY, Lee YL, Rupniewski I, Beausejour CM, Waite AJ, Wang NS, Kim KA, Gregory PD, Pabo CO, Rebar EJ. 2007. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat Biotechnol* 25: 778-85.
116. Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM. 2003. Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci U S A* 100: 5926-31. PMC ID: PMC156303.
117. Mitra RD, Shendure J, Olejnik J, Edyta Krzymanska O, Church GM. 2003. Fluorescent in situ sequencing on polymerase colonies. *Anal Biochem* 320: 55-65.
118. Moehle EA, Rock JM, Lee YL, Jouvenot Y, DeKaveler RC, Gregory PD, Urnov FD, Holmes MC. 2007. Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proc Natl Acad Sci U S A* 104: 3055-60. PMC ID: PMC1802009.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

119. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature* 430: 743-7.
120. Mrksich M. 2002. What can surface chemistry do for cell biology? *Curr Opin Chem Biol* 6: 794-7.
121. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Holmans P, Heward CB, Reiman EM, Stephan D, Hardy J. 2007. A survey of genetic human cortical gene expression. *Nat Genet* 39: 1494-9.
122. Myocardial Infarction Genetics C, Kathiresan S, Voight BF, Purcell S, Musunuru K, Ardissino D, Mannucci PM, Anand S, Engert JC, Samani NJ, Schunkert H, Erdmann J, Reilly MP, Rader DJ, Morgan T, Spertus JA, Stoll M, Girelli D, McKeown PP, Patterson CC, Siscovick DS, O'Donnell CJ, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Melander O, Altshuler D, Ardissino D, Merlini PA, Berzuini C, Bernardinelli L, Peyvandi F, Tubaro M, Celli P, Ferrario M, Fève R, Marziliano N, Casari G, Galli M, Ribichini F, Rossi M, Bernardi F, Zonzin P, Piazza A, Mannucci PM, Schwartz SM, Siscovick DS, Yee J, Friedlander Y, Elosua R, Marrugat J, Lucas G, Subirana I, Sala J, Ramos R, Kathiresan S, Meigs JB, Williams G, Nathan DM, MacRae CA, O'Donnell CJ, Salomaa V, Havulinna AS, Peltonen L, Melander O, Berglund G, Voight BF, Kathiresan S, Hirschhorn JN, Asselta R, Duga S, Spreafico M, Musunuru K, Daly MJ, Purcell S, Voight BF, Purcell S, Nemesh J, Korn JM, McCarroll SA, Schwartz SM, Yee J, Kathiresan S, Lucas G, Subirana I, Elosua R, Surti A, Guiducci C, Gianniny L, Mirel D, Parkin M, Burt N, Gabriel SB, Samani NJ, Thompson JR, Braund PS, Wright BJ, Balmforth AJ, Ball SG, Hall AS, Wellcome Trust Case Control C, Schunkert H, Erdmann J, Linsel-Nitschke P, Lieb W, Ziegler A, König I, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Schunkert H, Samani NJ, Erdmann J, Ouwehand W, Hengstenberg C, Deloukas P, Scholz M, Cambien F, Reilly MP, Li M, Chen Z, Wilensky R, Matthai W, Qasim A, Hakonarson HH, Devaney J, Burnett MS, Pichard AD, Kent KM, Sattler L, Lindsay JM, Waksman R, Epstein SE, Rader DJ, Scheffold T, Berger K, Stoll M, Häge A, Girelli D, Martinelli N, Olivieri O, Corrocher R, Morgan T, Spertus JA, McKeown P, Patterson CC, Schunkert H, Erdmann E, Linsel-Nitschke P, Lieb W, Ziegler A, König IR, Hengstenberg C, Fischer M, Stark K, Grosshennig A, Preuss M, Wichmann HE, Schreiber S, Holm H, Thorleifsson G, Thorsteinsdottir U, Stefansson K, Engert JC, Do R, Xie C, Anand S, Kathiresan S, Ardissino D, Mannucci PM, Siscovick D, O'Donnell CJ, Samani NJ, Melander O, Elosua R, Peltonen L, Salomaa V, Schwartz SM, Altshuler D. 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* 41: 334-41.
123. Nannya Y, Taura K, Kurokawa M, Chiba S, Ogawa S. 2007. Evaluation of genome-wide power of genetic association studies based on empirical data from the HapMap project. *Hum Mol Genet* 16: 2494-505.
124. Narayanan K, Warburton PE. 2003. DNA modification and functional delivery into human cells using *Escherichia coli* DH10B. *Nucleic Acids Res* 31: e51. PMC ID: PMC154239.
125. National Institutes of Health. 2009. Genotype-Tissue Expression Project (<http://nihroadmap.nih.gov/GTEX/>).
126. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324: 387-9.
127. Nolden L, Edenhofer F, Peitz M, Brüstle O. 2007. Stem cell engineering using transducible Cre recombinase. *Methods Mol Med* 140: 17-32.
128. Olsen PA, Randol M, Luna L, Brown T, Krauss S. 2005. Genomic sequence correction by single-stranded DNA oligonucleotides: role of DNA synthesis and chemical modifications of the oligonucleotide ends. *J Gene Med* 7: 1534-44.
129. Olsen PA, Solhaug A, Booth JA, Gelazauskaite M, Krauss S. 2009. Cellular responses to targeted genomic sequence modification using single-stranded oligonucleotides and zinc-finger nucleases. *DNA Repair (Amst)* 8: 298-308.
130. Pabo CO, Peisach E, Grant RA. 2001. Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem* 70: 313-40.
131. Pan X, Urban AE, Palejev D, Schulz V, Grubert F, Hu Y, Snyder M, Weissman SM. 2008. A procedure for highly specific, sensitive, and unbiased whole-genome amplification. *Proc Natl Acad Sci U S A* 105: 15499-504. PMC ID: PMC2563063.
132. Park CC, Ahn S, Bloom JS, Lin A, Wang RT, Wu T, Sekar A, Khan AH, Farr CJ, Lusk AJ, Leahy RM, Lange K, Smith DJ. 2008. Fine mapping of regulatory loci for mammalian gene expression using radiation hybrids. *Nat Genet* 40: 421-9.
133. Park IH, Zhao R, West JA, Yabuuchi A, Huo H, Ince TA, Lerou PH, Lensch MW, Daley GQ. 2008. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* 451: 141-6.
134. Pearson H. 2008. Protein engineering: The fate of fingers. *Nature* 455: 160-4.
135. Perez-Luz S, Abdulrazzak H, Grillo-Courvalin C, Huxley C. 2007. Factor VIII mRNA expression from a BAC carrying the intact locus made by homologous recombination. *Genomics* 90: 610-9.
136. Personal Genome Project. 2009. <http://www.personalgenomes.org/>.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

137. Plagnol V, Uz E, Wallace C, Stevens H, Clayton D, Ozelik T, Todd JA. 2008. Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS ONE* 3: e2966. PMC ID: PMC2494943.
138. Porreca GJ, Shendure J, Church GM. 2006. Polony DNA sequencing. *Curr Protoc Mol Biol* Chapter 7: Unit 7.8.
139. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J. 2007. Multiplex amplification of large sets of human exons. *Nat Methods* 4: 931-6.
140. Pruett-Miller SM, Connelly JP, Maeder ML, Joung JK, Porteus MH. 2008. Comparison of zinc finger nucleases for use in gene targeting in mammalian cells. *Mol Ther* 16: 707-17.
141. Pruett-Miller SM, Reading DW, Porter SN, Porteus MH. 2009. Attenuation of zinc finger nuclease toxicity by small-molecule regulation of protein levels. *PLoS Genet* 5: e1000376. PMC ID: PMC2633050.
142. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559-75. PMC ID: PMC1950838.
143. Radecke S, Radecke F, Peter I, Schwarz K. 2006. Physical incorporation of a single-stranded oligodeoxynucleotide during targeted repair of a human chromosomal locus. *J Gene Med* 8: 217-28.
144. Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273: 1516-7.
145. Roberts C, Chen CS, Mrksich M, Martichonok V, Ingber DE, Whitesides GM. 1998. Using Mixed Self-Assembled Monolayers Presenting RGD and (EG)3OH Groups To Characterize Long-Term Attachment of Bovine Capillary Endothelial Cells to Surfaces. *J Amer Chem Soc* 120: 6548-55.
146. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, International SNPMapWG. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928-33.
147. Sakamoto KM. 2005. Chimeric molecules to target proteins for ubiquitination and degradation. *Methods Enzymol* 399: 833-47.
148. Saxon E, Bertozzi CR. 2000. Cell surface engineering by a modified Staudinger reaction. *Science* 287: 2007-10.
149. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusk AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 6: e107. PMC ID: PMC2365981.
150. Schadt EE, Monks SA, Drake TA, Lusk AJ, Che N, Colnayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422: 297-302.
151. Sepp A, Choo Y. 2005. Cell-free selection of zinc finger DNA-binding proteins using in vitro compartmentalization. *J Mol Biol* 354: 212-9.
152. Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan JB, Hudson TJ. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet* 4: e1000006. PMC ID: PMC2265535.
153. Shao L, Feng W, Sun Y, Bai H, Liu J, Currie C, Kim J, Gama R, Wang Z, Qian Z, Liaw L, Wu WS. 2009. Generation of iPS cells using defined factors linked via the self-cleaving 2A sequences in a single open reading frame. *Cell Res* 19: 296-306.
154. Shendure J, Mitra RD, Varma C, Church GM. 2004. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* 5: 335-44.
155. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309: 1728-32.
156. Shendure JA, Porreca GJ, Church GM. 2008. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol* Chapter 7: Unit 7.1.
157. Shi Y, Despons C, Do JT, Hahm HS, Scholer HR, Ding S. 2008. Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. *Cell Stem Cell* 3: 568-74.
158. Shin J, Kayser SR, Langae TY. 2009. Pharmacogenetics: from discovery to patient care. *Am J Health Syst Pharm* 66: 625-37.
159. Smith EN, Kruglyak L. 2008. Gene-environment interaction in yeast gene expression. *PLoS Biol* 6: e83. PMC ID: PMC2292755.
160. Smith J, Modrich P. 1997. Removal of polymerase-produced mutant sequences from PCR products. *Proc Natl Acad Sci U S A* 94: 6847-50. PMC ID: PMC21247.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

161. Srivastava M, Hsieh S, Grinberg A, Williams-Simons L, Huang SP, Pfeifer K. 2000. H19 and Igf2 monoallelic expression is regulated in two distinct ways by a shared cis acting regulatory region upstream of H19. *Genes Dev* 14: 1186-95. PMC ID: PMC316622.
162. Stark JM, Pierce AJ, Oh J, Pastink A, Jasin M. 2004. Genetic steps of mammalian homologous repair with distinct mutagenic consequences. *Mol Cell Biol* 24: 9305-16. PMC ID: PMC522275.
163. Stougaard M, Lohmann JS, Zajac M, Hamilton-Dutoit S, Koch J. 2007. In situ detection of non-polyadenylated RNA molecules using Turtle Probes and target primed rolling circle PRINS. *BMC Biotechnol* 7: 69. PMC ID: PMC2203993.
164. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315: 848-53.
165. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P, Dermitzakis ET. 2007. Population genomics of human gene expression. *Nat Genet* 39: 1217-24.
166. Suzuki T. 2008. Targeted gene modification by oligonucleotides and small DNA fragments in eukaryotes. *Front Biosci* 13: 737-44.
167. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131: 861-72.
168. Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126: 663-76.
169. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6: 377-82.
170. Tao H, Cox DR, Frazer KA. 2006. Allele-specific KRT1 expression is a complex trait. *PLoS Genet* 2: e93. PMC ID: PMC1475705.
171. Tatum EL, Lederberg J. 1947. Gene Recombination in the Bacterium *Escherichia coli*. *J Bacteriol* 53: 673-84. PMC ID: PMC518375.
172. Terry R, Porreca G, McCarthy K, Church GM. 2008. Polonator Instrument <http://www.polonator.org>
173. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, Stacey SN, Bergthorsson JT, Thorlacius S, Gudmundsson J, Jonsson T, Jakobsdottir M, Saemundsdottir J, Olafsdottir O, Gudmundsson LJ, Bjornsdottir G, Kristjansson K, Skuladottir H, Isaksson HJ, Gudbjartsson T, Jones GT, Mueller T, Gottsater A, Flex A, Aben KK, de Vegt F, Mulders PF, Isla D, Vidal MJ, Asin L, Saez B, Murillo L, Blondal T, Kolbeinsson H, Stefansson JG, Hansdottir I, Runarsdottir V, Pola R, Lindblad B, van Rij AM, Dieplinger B, Haltmayer M, Mayordomo JI, Klemeney LA, Matthiasson SE, Oskarsson H, Tyrfinsson T, Gudbjartsson DF, Gulcher JR, Jonsson S, Thorsteinsdottir U, Kong A, Stefansson K. 2008. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 452: 638-42.
174. Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, Church G. 2004. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432: 1050-4.
175. Townsend JA, Wright DA, Winfrey RJ, Fu F, Maeder ML, Joung JK, Voytas DF. 2009. High Frequency Modification of Plant Genes Using Engineered Zinc Finger Nucleases. *Nature*: advanced on-line print: April 29, 2009.
176. Turner DJ, Shendure J, Porreca G, Church G, Green P, Tyler-Smith C, Hurles ME. 2006. Assaying chromosomal inversions by single-molecule haplotyping. *Nat Methods* 3: 439-45.
177. University of Washington (Seattle). 2009. [genetests.org](http://www.genetests.org) (<http://www.genetests.org/>). pp. Quote from web page (03/24/09): 475 GeneReviews, 1,158 Clinics, 607 Laboratories testing for 1,707 Diseases: 1,422 Clinical, 285 Research
178. Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC. 2005. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* 435: 646-51.
179. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* 5: 829-34.
180. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabriellian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S,

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

- Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooshep S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Foster C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. The sequence of the human genome. *Science* 291: 1304-51.
181. Vigneault F, Sismour AM, Church GM. 2008. Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation. *Nat Methods* 5: 777-9.
 182. Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* 9: 255-66.
 183. Wade-Martins R, Smith ER, Tyminski E, Chiocci EA, Saeki Y. 2001. An infectious transfer and expression system for genomic DNA loci in human and mouse cells. *Nat Biotechnol* 19: 1067-70.
 184. Wang HH, Isaacs FJ, Forest CR, Sun ZZ, Xu G, Church GM. 2009. Programming cells by multiplex genome engineering and accelerated evolution *Nature* in press.
 185. Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57-63.
 186. Waters VL. 2001. Conjugation between bacterial and mammalian cells. *Nat Genet* 29: 375-6.
 187. Watkins H, Farrall M. 2006. Genetic susceptibility to coronary artery disease: from promise to progress. *Nat Rev Genet* 7: 163-73.
 188. Wellcome Trust Case Control C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-78.
 189. Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, Hamalainen R, Cowling R, Wang W, Liu P, Gertsenstein M, Kaji K, Sung HK, Nagy A. 2009. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*.
 190. Wu J, Zhang S, Meng Q, Cao H, Li Z, Li X, Shi S, Kim DH, Bi L, Turro NJ, Ju J. 2007. 3'-O-modified nucleotides as reversible terminators for pyrosequencing. *Proc Natl Acad Sci U S A* 104: 16462-7. PMC ID: PMC2034218.
 191. Yanez RJ, Porter AC. 1998. Therapeutic gene targeting. *Gene Ther* 5: 149-59.
 192. Ying QL, Wray J, Nichols J, Battle-Morera L, Doble B, Woodgett J, Cohen P, Smith A. 2008. The ground state of embryonic stem cell self-renewal. *Nature* 453: 519-23.
 193. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL. 2000. An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc Natl Acad Sci U S A* 97: 5978-83. PMC ID: PMC18544.
 194. Yu J, Hu K, Smuga-Otto K, Tian S, Stewart R, Slukvin, II, Thomson JA. 2009. Human induced pluripotent stem cells free of vector and transgene sequences. *Science* 324: 797-801.
 195. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, Slukvin, II, Thomson JA. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318: 1917-20.
 196. Zahnd C, Amstutz P, Pluckthun A. 2007. Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. *Nat Methods* 4: 269-79.
 197. Zaranek AW, Clegg W, Vandewege, Church GM. 2008. *Free Factories: Unified Infrastructure for Data Intensive Web Services* Presented at USENIX Annual Technical Conference, Boston, MA
 198. Zhang D, Baek SH, Ho A, Lee H, Jeong YS, Kim K. 2004. Targeted degradation of proteins by small molecules: a novel tool for functional proteomics. *Comb Chem High Throughput Screen* 7: 689-97.
 199. Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee J, Aach J, Leproust E, Eggan K, Church GM. 2009. Digital RNA Allelotyping Reveals Tissue-specific and Allele-specific Gene Expression in Human (submitted to *Nature Methods*).
 200. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM. 2006. Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* 24: 680-6.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

201. Zhang K, Zhu J, Shendure J, Porreca GJ, Aach JD, Mitra RD, Church GM. 2006. Long-range polony haplotyping of individual human chromosome molecules. *Nat Genet* 38: 382-7.
202. Zhou Y, Calciano M, Hamann S, Leamon JH, Strugnell T, Christian MW, Lizardi PM. 2001. In situ detection of messenger RNA using digoxigenin-labeled oligonucleotides and rolling circle amplification. *Exp Mol Pathol* 70: 281-8.
203. Zhu J, Shendure J, Mitra RD, Church GM. 2003. Single molecule profiling of alternative pre-mRNA splicing. *Science* 301: 836-8.
204. Zondervan KT, Cardon LR. 2004. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 5: 89-100.
205. Zondervan KT, Cardon LR. 2007. Designing candidate gene and genome-wide case-control association studies. *Nat Protoc* 2: 2492-501.
206. Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9: 133-48. PMC ID: PMC326673.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Management and Organization

- A. Overview
- B. Organizational Structure
- C. Key investigators
- D. Resources
- E. Administrative and financial structure
- F. Monitoring of progress
- G. Conflict of Interest Policy
- H. Timeline and milestones

A. Overview.

The biological focus of our proposed Center for the Causal Transcriptional Consequences of Human Genetic Variation (CTCHGV) is to identify natural human genetic variations that cause differential transcription in cis genes by directly modifying the DNA at sites of variation in human cells and observing the effects on transcriptional levels. To this end, CTCHGV will develop: (1) innovative and scalable methods for precisely engineering DNA in human cells, (2) new methods for synthesizing thousands of zinc finger nuclease (ZFN) proteins that will enable this engineering, (3) technology for generating and using human induced Pluripotent Stem cells (iPS) to track the impact of the engineered changes in many human cell types, and (4) transcriptome level RNA expression assays that will operate in single human cells. CTCHGV methods will be extremely timely in that they promise to extend beyond and refine the results obtained by Genome Wide Association Studies (GWAS) over the past several years, which are good at identifying variable loci that are associated with phenotypes and disorders, but cannot generally identify which among many correlated variations at these loci are responsible for them. CTCHGV's development of methods that directly identify causality will therefore have broad impact on biomedical research generally and will be enabling for personalized medicine, while the engineering and synthesis components will have very direct impacts on gene therapy and synthetic biology.

CTCHGV will be led by Professor George M. Church of Harvard Medical School, who has already led a prior Molecular and Genomic Imaging CEGS (MGIC) with a solid record of productivity and accomplishment. MGIC generated 44 manuscripts in years 2-5 that were published in peer reviewed journals (see References), and had close collaborative relationships with 23 companies concerning CEGS-supported sequencing technology and applications (see Data and Materials Dissemination Plan). CTCHGV's objectives and orientation are significantly different than MGIC's and CTCHGV is being proposed not as a renewal of MGIC but as a new and independent Center. Accordingly, although CTCHGV and MGIC have some membership in common—namely, Professors Church and Kun Zhang (UCSD)—its other membership has changed: the new co-Investigators are Professors George Q. Daley and J. Keith Joung (both from Harvard Medical School).

Integration, communication, and impact: The MGIC CEGS found frequent communication between co-Investigators and director Professor Church to be the most effective method of ensuring integration and alignment of goals, and this will continue in CTCHGV. Except for the Zhang Lab in UCSD, all of CTCHGV will be in the Boston area, making communication easy. The Zhang Lab is remote from Boston, but Dr. Zhang and Professor Church have a close relationship based on the Dr. Zhang's time as a post-doc in the Church Lab and current collaboration in the context of an NHLBI grant (HLB08-004). The Boston focus will allow CTCHGV to interact easily with the region's concentrated cluster of major and influential research institutions (Broad Institute, Whitehead Institute, MIT, Boston University, U. Mass), assuring both access to a rich base of research resources and the ability to broadly disseminate CTCHGV technology development to the wider research community.

As a matter of principle, Professor Church strongly believes in open dissemination of knowledge and technology, and is therefore committed to making CTCHGV innovations available to the larger research community: both directly through tools and methods for immediate use by individual researchers, and by technology transfer to industry, whereby companies incorporate the innovations into their products. Low cost, scalability, open methods and protocols, and quantitative and objective reliability assessments are high

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

priorities for all CTCHGV technology development and are integrated into all Aims. Moreover, all CTCHGV technologies will be packaged with demonstrations of how they can be used to address important biological research questions, thus providing the research community with models they can follow to develop further applications.

Achievability of CTCHGV goals: CTCHGV features leading researchers in the relevant science and technology – Professor Church in synthetic biology and high throughput assay development, Professor Daley in stem cell biology, and Professor Joung in ZFN development and optimization. CTCHGV also features many collaborators who will bring in other important resources and areas of expertise, such as Professors David Altschuler, Steven McCarroll, and Robert Plenge for GWAS expertise, Professor Steven Elledge for genome-wide screening in human cells, and Complete Genomics, Inc. for sequencing capability (see Letters of Support). Careful attention has been given in the CTCHGV proposal to setting an achievable but ambitious scope. Formally, CTCHGV's Aims are explicitly identified as the *development and demonstration* of innovative and highly scalable methods and not as large-scale and systematic applications of these methods, which would exceed CTCHGV resources and is considered out of scope. However, CTCHGV has set ambitious numerical targets for the demonstration of its methods – e.g., the analysis of 1000 human genes for cis causation (Aim 1) – with the understanding that these targets will be reviewed at the end of the second year of the Center and renegotiated as needed with NIH in the light of both CTCHGV progress and the progress of the field as a whole (see Research Design and Methods, Overview). Among the reasons why we feel ambitious targets are important are: (i) we believe strongly that the technologies under consideration shouldn't be underestimated and (ii) only with ambitious targets can we properly test *scalability*. Finally, Professor Church's many years of experience directing large Centers and research projects is itself a success factor for CTCHGV.

B. Organizational Structure

Professor Church directs the Center, with assistance from John Aach (Lecturer) and Yveta Masarova (Admin Assistant). The Church Lab will work on all Aims, and all other Labs will work on subsets of the Aims under Professor Church's direction. The Joung Lab specializes in ZFN development and optimization and will work on Aims 1.1 and 4.2, which develop ZFNs for Aim 1's analysis of cis gene regulation, and develop novel technology for ZFN optimization, respectively. The Daley Lab specializes in human stem cells and iPS, and will work on Aim 2, which focuses on tracking of Aim 1-characterized regulation in multiple human cell types derived from genetically engineered human iPS. The Zhang Lab develops high-throughput methods for characterization of allele-specific expression, long range haplotyping, and targeted sequencing, all of which are components of Aim 1. CTCHGV will set up a board of Scientific Advisors to provide ongoing advice and input to Professor Church (see F. Monitoring of Progress below).

C. Key investigators.

George Church, Ph.D. will direct the proposed center. He is Professor of Genetics at Harvard Medical School and Director of the Lipper Center for Computational Genetics. His Ph.D. work from Harvard in Biochemistry & Molecular Biology with Wally Gilbert and as Scientist at Biogen included the first direct genomic sequencing method in 1984 and the co-initiation of the Human Genome Project. As Life Sciences Research Fellow at UCSF with Gail Martin he was among the first to work on embryonic stem cells (in 1985). He invented the broadly-applied concepts of molecular multiplexing and tags, and applications of array DNA synthesizers. Technology transfer of automated sequencing and software to Genome Therapeutics Corp. resulted in the first commercial genome sequence (the human pathogen, *H. pylori*, 1994). This multiplex solid-phase sequencing evolved into polonies in 1999, ABI-SOLiD in 2005, and open-source Polonator.org in 2007. He has served in advisory roles for 12 journals, 5 granting agencies, and 23 biotech companies (recently founding Knome and LS9). Current research focuses on cost-effective next generation sequencing, targeted sequencing, synthetic biology, and scalable personal genomics. John Aach, Ph.D. is a Lecturer in the Department of Genetics who has been a computational biologist in the Church Lab since 1996. His research has focused on development of computational and statistical methods and applications for multiple high-throughput technologies, including sequence, expression, images, and mathematical models. He will assist Professor Church in overseeing Center activities as well as work on computational aspects of Center research Aims.

George Daley, M.D., Ph.D. is an Associate Professor of Pediatrics in the Division of

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Hematology/Oncology at the Children's Hospital and Dana Farber Cancer Institute, an Associate Professor of Biological Chemistry and Molecular Pharmacology at Harvard Medical School, and an HHMI investigator. Dr. Daley is an internationally recognized stem cell researcher whose laboratory has developed leading methods for generating induced Pluripotent Stem (iPS) cells from fibroblasts from human skin biopsies of healthy and diseased individuals. Dr. Daley received a Ph.D. in biology from MIT (1989) and the M.D. *summa cum laude* from Harvard Medical School (1991). Dr. Daley is Board Certified in Internal Medicine and Hematology and is currently a staff physician in Hematology/Oncology at the Children's Hospital, the Dana Farber Cancer Institute, and the Brigham and Women's Hospital in Boston. He has been elected to the American Society for Clinical Investigation and has received major awards from the American Philosophical Society, the Society for Pediatric Research, the Burroughs Wellcome Fund, and the Leukemia and Lymphoma Society of America. He received the inaugural NIH Director's Pioneer Award, a five-year unrestricted grant to pursue highly innovative research. Dr. Daley's work focuses on functional hematopoietic and germ cell elements from ES cells, and the genetic mechanisms that predispose to malignancy. Dr. Daley's lab was one of the first three world-wide to derive human iPS cells, and the first to produce a repository of patient-specific iPS cells (from 10 different disease conditions).

J. Keith Joung, M.D., Ph.D. is an Associate Professor of Pathology at Harvard Medical School, and Associate Chief of Pathology for Research and Director of the Molecular Pathology Unit at Massachusetts General Hospital (MGH). He is also a member of the Center for Cancer Research and of the Center for Computational and Integrative Biology at MGH. He received his M.D. and Ph.D. (in Genetics) from Harvard Medical School in 1996. His Ph.D. work (with Ann Hochschild) focused on mechanisms of transcriptional regulation in prokaryotes. He completed residency training in Clinical Pathology at Massachusetts General Hospital in 1999 and a post-doctoral research fellowship in Carl Pabo's lab at the Massachusetts Institute of Technology in 2001. Dr. Joung's research interests include understanding how Cys₂His₂ zinc fingers, the most common domain encoded in the human genome, mediate specific protein-DNA and protein-protein interactions. In addition, Dr. Joung's lab uses a combination of directed randomization and bacterial cell-based selection methods to engineer artificial "designer" zinc finger domains with desired DNA-binding specificities. Designer zinc fingers hold promise as a tool for altering any genomic locus of interest and have tremendous potential in both research and gene therapy applications. Dr. Joung is the leader and co-founder of the Zinc Finger Consortium (<http://www.zincfingers.org/>), which was established to ensure and to promote continued research and development of engineered zinc finger technology. The Consortium is committed to developing a zinc finger engineering platform that is robust, user-friendly, and freely available to the academic scientific community.

Kun Zhang, Ph.D. is an Assistant Professor of Bioengineering at the Jacobs School of Engineering at University of California in San Diego. His recent research has focused on developing assays for single DNA molecules and applying these methods to human genetic and environmental genomic studies. Prior to joining UCSD, Professor Zhang was a post-doc in the Church Lab where he developed Multiplex Amplification from Single Molecules (MASMO), an efficient multi-locus haplotyping on human chromosome molecules which has many applications including linkage disequilibrium analyses in large populations and characterization or mapping of recombination/translocation crossovers in mammalian cells. In collaboration with James Gusella's lab at Massachusetts General Hospital, Zhang and his colleagues used MASMO to map breakpoints of balanced chromosome translocation in patients with a variety of developmental disorders. Professor Zhang also developed methods to amplify DNA from single cells or chromosomes for genome sequencing, which can be used to completely sequence unculturable organisms from single cells and is also being applied to sequence individual cancer chromosomes to characterize the genome anatomy of cancer genome at unprecedented resolution.

D. Resources

The proposed center will include Harvard, Harvard Medical School (and its hospital affiliates), and UCSD. There will be no difficulties for the sharing of equipment between the Harvard sites. As noted above, these CTCHGV labs will also have access to many facilities and resources in the Boston area, including the Broad Institute. Locally, the Harvard Medical School Biopolymers Facility has available an Illumina Genome Analyzer II, while the Broad Institute recently acquired 22 of them. The Church Lab has had access to a 454 FLX sequencer through a Private Source grant which is focused on sequencing the repertoire of VDJ elements in

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

B cell antibodies and T cell receptors in human subjects. In connection with single molecule sequencing, as described in section 5.3.1, Harvard has recently acquired a Heliscope single molecule sequencing device, and Professor Church is on the Scientific Advisory Board of Helicos where he can obtain access to expertise. Additionally, Complete Genomics, Inc. is offering sequencing support (see Letters of Support). Computational resources are available at Harvard Medical School through the HMS Research Information Technology Group's shared "orchestra" cluster, which at this writing has 179 compute nodes, 810 processor cores, 1.66TB aggregate memory (4-32GB per core), and 23.78 TB of local disk capacity. To ensure dedicated vs. shared access to processing capability, we have budgeted for 10 compute nodes to be added to this cluster for CTCHGV processing (see Budget Justification, yr 1, Specialized Equipment). Through Harvard, MIT, and the Broad Institute, Professor Church has access to other expertise and resources for CTCHGV projects, including access to materials and expertise needed for RNAi and overexpression screens for optimization of homologous recombination in Aims 1 and 4, as well as expertise on GWAS for Aim 1.4 (see Letters of Support from David Altshuler, Steve McCarroll, Robert Plenge, and Steve Elledge, among others). Microfluidics facilities are available locally at the Harvard Medical School (<http://microfluidics.hms.harvard.edu/>).

E. Administrative and financial structure.

Professor Church will oversee the administrative and financial aspects of the proposed center. He will be assisted by the administrative and financial officers of the Dept. of Genetics at HMS working in conjunction with their counterparts at Harvard affiliates and UCSD. While each key investigator has submitted a detailed first year budget and a five year plan, it is likely that some adjustments will become necessary in order to keep our priorities in alignment with discoveries and cost-saving technological advancements. Any adjustments will be made by the Director, in close consultation with key investigators and with prior approval from our financial officers and the NIH.

F. Monitoring of progress.

As noted above, Professor Church will have frequent interactions (at least one per month by email, phone, or in person) with all key investigators in which he will monitor progress and align directions of all teams in CTCHGV. Since most activity of CTCHGV will be in the Boston area, he will encourage teams to send representatives to each other's team meetings, a good way of communicating progress details across the Center. As Dr. Zhang at UCSD is geographically remote from Boston, Professor Church will primarily keep in touch by email and phone on a weekly basis and will commit to meeting personally at least once a year.

The yearly progress report to NIH will set the occasion for yearly consultation with all key investigators on overall progress and direction of the CTCHGV. A yearly meeting of key investigators will be arranged for discussion and reprioritization of CTCHGV directions, with Dr. Zhang participating by conference call if he cannot be physically present. Furthermore, CTCHGV has defined a set of CTCHGV Scientific Advisors with whom Professor Church will consult for input and advice on an on-going basis concerning CTCHGV progress and direction. Additionally, the annual NIH progress report will be circulated to these Scientific Advisors as a way of stimulating discussion and oversight on an annual basis. The Scientific Advisors will include:

- Professor Stephen Elledge (Harvard Medical School, HHMI)
- Professor David Altshuler (Broad Institute, Massachusetts General Hospital)
- Professor Robert Plenge, M.D., Ph. D. (Harvard Medical School, Brigham and Women's Hospital)

G. Conflict of Interest Policy

All investigators will be expected to adhere to institutional Conflict of Interest policies.

H. Timeline and milestones.

The timeline below provides our initial estimate of when certain milestones will be achieved. It integrates the long-term and intermediate goals that are described at the end of every Aim described in the Research Design plan (section 5). In on-going and yearly assessments CTCHGV will consider whether the timeline or overall goals need to be adjusted. Additionally, in Research Design section 5 (Overview), we have specifically programmed consultation and possible renegotiation with NIH of CTCHGV targets at the end of

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

year 2 of the Center in the light of progress, both at CTCHGV and the field as a whole. The project development schedule below is meant to be flexible rather than constraining.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

CTCHGV sub-Aim

Aim	CTCHGV Year				
	0	1	2	3	4
1.1	ZNF development & set-up				
1.1	MAGE-BAC development & set-up				
1.1	MAGE-human development & set-up				
1.1	genes 1-50: altered cis region populations		genes 51-1000: altered cis region populations		
1.2	genes 1-50: identify causative cis variants		genes 51-1000: identify causative cis variants		
1.2	develop 1-cell ASE/genotyping method		use 1-cell ASE/genotyping method		
1.3	assess splicing, CNV, epigenetics, &c in CTCHGV original and altered cell lines				
1.4			assess relationship between GWAS and CTCHGV methods		
2.1.2.2	develop methods for creating / maintaining IPS with altered cis variations		explore ASE, altered cis variants in 50 genes in 3 IPS-derived cell types		
2.3	create IPS "marked allele" cell lines: 5 genes		create IPS "marked allele" cell lines: 50 genes in 3 subject cell lines		
3.1.3.2	develop and evaluate single cell sequencing approaches		apply single cell sequencing to differentiating IPS: human primary skin+IPS fibroblasts		
3.1.3.2	single cell sequencing of 100 transcripts/cell		single cell sequencing of 1000 transcripts/cell		
4.1	initial demonstration of integrated DNA sequencing/synthesis platform		1000 ZFN created using integrated DNA sequencing/synthesis		
4.2	all zinc finger pools completed		10000 zinc finger triplexes optimized via ribosome display		
4.3	develop cell handling for Aims 1-2		demonstrate cell sorting by morphology		

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Training Plan

- A. Overview
- B. CTCHGV researcher training in genomics
- C. CTCHGV education for non-CTCHGV researchers
- D. CTCHGV genomics education for the public
- E. CTCHGV consultant for training

A. Overview. As in our prior MGI CEGS, our primary goal will be to train students at all levels (summer interns and undergraduates, grad students, post-docs) for careers in genomics and related science by having them take part in CTCHGV research under the tutelage of CTCHGV investigators and other experts. We had great success with such training in MGIC: 22 MGIC trainees went on to positions in science and 14 are either still in MGIC labs or are otherwise continuing their education (see Table TP-1). In addition to taking part in CTCHGV, we will also promote cross-training and disseminate information in genomics to scientists outside of CTCHGV, as we did in MGIC (see B below). Second, through talks, seminars, and classroom teaching, we will educate the wider research community about CTCHGV technology, its applications, and implications.

Table TP-1: Researcher training under MGIC

MGIC Trainees who went on to careers in genomics, science, medicine

<i>Trainee</i>	<i>Position</i>	<i>MGIC Lab</i>
Vasudeo Badarinarayana, PhD	Senior Scientist, Sanofi-Aventis	Church
Samuel Boutin, PhD	Veterinarian Scientist, Taconic, Germantown, NY	Sherley
Joao Pedro de Magalhaes, PhD	Assistant Professor, University of Liverpool	Church
Patrik d'Haeseleer, PhD	Scientist, Lawrence Livermore National Lab	Church
Aimee Dudley, PhD	Assistant Professor, Institute of Systems Biology	Church
Gary Gao, PhD	Assistant Professor, Virginia Commonwealth University	Church
Craig Forest, PhD	Assistant Professor, Georgia Institute of Technology,	Church
Yonatan Grad, MD/PhD	Resident, Brigham & Womens Hospital, Boston, MA	Church
Matt Hofelder, technician	Technical Sales at Sigma → Applied Biosystems	Mitra
Janice A. Lansita, PhD	Toxicologist, Biogen-IDEC, Cambridge, MA	Sherley
Debi Mitra, MD (MGIC summer employee)	Resident, Vanderbilt University	Mitra
Min Soo Noh, PhD	Scientist, AMOREPACIFIC Corporation, South Korea	Sherley
Dana Pe'er, PhD	Assistant Professor, Columbia University	Church
Gregory Porreca, PhD	Senior Scientist, Good Start Genetics, Inc.	Church
Nick Reppas, PhD	Research Scientist, Joule Biotechnologies, Inc.	Church
Jay Shendure, PhD	Assistant Professor, University of Washington	Church
Martin Steffen, MD/PhD	Assistant Professor, Boston University School of Medicine	Church
Chris Varma, PhD	Partner, Flagship Ventures, Cambridge, MA	Church
Dennis Vitkup, PhD	Assistant Professor, Columbia University	Church
Kun Zhang, PhD	Assistant Professor, University of California at San Diego	Church
Jun Zhu, PhD	Assistant Professor, Duke University	Church
Zhou Zhu, PhD	Senior Scientist, Pfizer Labs, La Jolla	Church

Current MGIC Trainees or former trainees continuing their education

Adnan Derti, PhD		Post-doc, Roth Lab, Harvard Medical School	Church
Personal Info	summer intern	Completing PhD, cell migration biological engineering	Sherley
	summer intern	Completing undergraduate degree, College of St. Scholastica	Sherley
Jin Billy Li, PhD		Post-doc, Church Lab	Church
German Leparo, PhD		Post-doc, Kreil Lab, Boku University, Vienna	Mitra

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Personal Info	undergraduate	Medical school matriculant	Mitra
Katherine Montero, technician		Applying to business school	Church
Yasmine Ndassa, grad student		Grad student, Marto Lab, Dana Farber	Church
Amy Nichols, grad student		Completing PhD, proteomics	Sherley
Personal Info	summer intern	Medical school matriculant	Sherley
Madeleine Price Ball, grad student		Grad student, Church Lab	Church
Abraham Rosenbaum, grad student		Grad student, Church Lab	Church
Sarah Vassallo, technician		technician, Church Lab	Church
Alexander Zaranek		Post-doc, Church Lab	Church

B. CTCHGV researcher training in genomics. As noted above, CTCHGV will provide in-depth training in genomics and related science by having students at all levels work with CTCHGV PIs and experts in their Labs. We will also promote cross-training by encouraging students to attend lab meetings in other CTCHGV labs and will also encourage them to actually work in these labs where protracted contact is important. Under MGIC, versions of this arrangement were used when Samuel Boutin (Sherley Lab) worked for several months in the Church Lab, and Jeremy Edwards (UNM) came for an intensive week, to learn polony techniques at the bench in order to take them back to their own labs. Within CTCHGV as proposed, most labs are in the Boston area making these arrangements quite feasible; and while it will be more difficult for members of the Zhang Lab (USCD) to spend time on other CTCHGV labs, Zhang's close knowledge of methods in the Church Lab will also make them less necessary.

Because CTCHGV will work with samples from human subjects, CTCHGV will require that all of its researchers working with human subject samples and data to take institutional training on human subjects protection, e.g., for Harvard affiliates, Harvard offers a web-based training and certification program (HETHR) on human subjects protection.

C. CTCHGV education for non-CTCHGV researchers. CTCHGV will help educate the research community generally by publishing papers and protocols (for MGIC's high productivity in this regard, see the Management and Organization Plan and also References) and also by giving talks in classes and conferences. CTCHGV's concentrated presence in the Boston area will spread word of CTCHGV developments in the influential Boston biomedical research community. Notably, Professor Church gives talks frequently both in and outside of the Boston area, and is a participant and speaker at many seminars in the Boston area that meet regularly. He also directs and is the sole instructor of the main course on genomics and computational biology at HMS, Harvard College, Harvard Extension School, and MIT, and also has participated in team-taught courses, including MIT, HSPH, JFK School of Government, Harvard Business School, and Boston University College of Engineering. His primary course is also accessible through the Internet and, in this way, he reaches students internationally (<http://www.courses.fas.harvard.edu/~bphys101/>). The course has also been made available via MIT OpenCourseWare (<http://ocw.mit.edu/index.html>). Professor Church will commit to giving four talks a year that focus on CTCHGV technology and its implications. In addition, the investigators involved in CTCHGV have regular teaching responsibilities. Professor Church will give investigators his own class notes and presentations that deal with CTCHGV topics to help provide material and focus to their own talks and teaching session on these topics, and he will encourage all investigators to cover these topics and to share their notes and materials with each other.

D. CTCHGV genomics education for the public Professor Church, the proposed director of CTCHGV, is also the director of the Personal Genome Project (PGP, <http://www.personalgenomes.org/>), which maintains web-based enrollment and training tools by which the public can obtain information on and improve their knowledge of genetics. CTCHGV will use this connection to establish cross-links between CTCHGV and PGP web sites and will develop web based resources to help the public understand the importance of CTCHGV developments. Professor Church very frequently gives interviews on trends and developments in genetics to reporters for print, radio, and television media, and he will also use these connections and forums to improve public awareness of CTCHGV developments.

E. CTCHGV consultant for training As the CTCHGV develops methods that will identify and characterize cause-effect relationships between human genome sequence variation and transcriptional networks, we expect opportunities to arise for outside investigators to utilize these techniques to explore

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

human disease associations. The CTCHGV has asked Robert C. Green, MD, MPH to assist the Center with evaluating these opportunities and consult to the Center on any research involving human subjects. Dr. Green is an established researcher in genetic epidemiology, translational genomics and ELSI issues and will serve as a consultant to the Center. In addition to having received continuous NIH funding for 21 years, Dr. Green has recently been awarded a K24 clinical research mentoring award from NIH, and has extensive experience mentoring minority trainees and (in his research studies) recruiting minority participants. With this background, he will also assist in the multi-disciplinary teaching of pre-doctoral and post-doctoral trainees, as well as junior faculty, by providing structured lectures at Center gatherings on clinical research issues and by providing individual mentoring to Center trainees who are interested in, or collaborating with, clinical investigation.

Minority Action Plan

A. Summary

B. CTCHGV MAP Program Description

C. CTCHGV MAP Management

D. MGI CEGS MAP Program Progress

A. Summary. We fully recognize the need for a pro-active stance in addressing the under-representation of minorities in science and ELSI research. We will build upon our extensive track record for implementing successful minority recruiting and training programs and leverage existing infrastructures, associated with Harvard Medical School and other institutions affiliated with our center, to increase the intellectual strength of genomic science and provide high impact results for the professional careers of trainees.

The CTCHGV will benefit from our five years of experience with managing successful minority training as part of our previous MGIC CEGS, which has enrolled and mentored 16 under-represented minority trainees since 2005. Historically, our training programs with the most demonstrated success and the highest impact on the professional development of trainees are at the undergraduate and postdoctoral levels (See Program Successes below). We will focus on these programs for CTCHGV, as we believe they will continue to provide the most valuable training opportunities we can offer for the professional development of minorities seeking careers in genomics and ELSI research.

Over five years, CTCHGV MAP aims to recruit, train, and support the activities of 25 undergraduates, 5 post doctoral fellows, and up to 5 postbacs. This program will enable trainees to participate in research activities with multi-investigator, interdisciplinary teams that will develop innovative genomic approaches to address biological problems.

B. CTCHGV MAP Program Description. The CTCHGV training program will focus on activities at three career levels: training postdocs for professional advancement, postbacs for placement in graduate programs, and undergraduate students for research experience and preparation for graduate school. This training program will enable trainees to engage in research activities in an unparalleled research environment that combines highly interdisciplinary projects in the genomic sciences with leaders in the field. The center includes investigators from Harvard Medical School and affiliates, and UCSD. We will utilize existing minority programs at these institutions. Additionally, CTCHGV has included as a consultant Dr. Robert Green (see Training Plan) with considerable experience in minority recruitment and training for biomedicine who has access to Boston University programs and resources set up for these goals. By this means, we will significantly broaden our recruitment efforts.

B.1 Recruitment. Representatives from the center will attend the two largest annual meetings in the New England region for recruiting minorities for genomic and biomedical science research activities: the New England Science Symposium (NEST) and the Biomedical Sciences Careers Program (BSCP) conference. Both occur in the Spring of every year. In the past, these meetings have attracted over 300 underrepresented minority postdocs, graduate, medical, and undergraduate students. Existing minority training programs (see section B.3 below) will be leveraged for additional recruitment and outreach activities at the undergraduate level.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

B.2 Postdoctoral Training Program. The CTCHGV will engage postdoctoral fellows in a highly interdisciplinary research environment that combines genomic, computational, and technology-development with traditional biological sciences. We will support "bridging" research positions that will permit post-doc trainees to work closely with multiple key investigators of the center. This approach will provide our trainees with the experiences necessary to prepare for leadership in the genomic, computational, and biological sciences.

Milestones: CTCHGV CEGS MAP will support two postdoctoral fellows per year (See Table MAP-1). To facilitate the transition from one educational level to another, postdocs will be encouraged in their second year to participate in academic enhancement activities, such as the submission of applications for young investigator awards F32, K99, and R00.

Table MAP-1: CTCHGV CEGS MAP Postdoc Training Program					
	2009	2010	2011	2012	2013
Dr. Adeyemi Adesokan	2 nd year				
Postdoc #2	1 st year	2 nd year			
Postdoc #3		1 st year	2 nd year		
Postdoc #4			1 st year	2 nd year	
Postdoc #5				1 st year	2 nd year

B.3. Undergraduate Summer Research Experience. We firmly believe that the impact of an undergraduate research experience can be a turning point for convincing young investigators that research is a very rewarding and feasible career choice. Interaction between undergraduate students and center mentors will encourage the pursuit of graduate degrees, lead to publications, prepare students for entry into graduate programs, and facilitate the ability of undergraduates to obtain letters of recommendation.

Milestones: We will recruit and provide funding for 6 underrepresented minority undergraduate students every year. To achieve these goals, we will work with the existing programs at Harvard Medical School and other institutions affiliated with the center. We have identified 5 minority programs at 4 institutions to collaborate with for the recruitment and enrollment of undergraduates in our Summer Research Experience program:

SHURP. Summer Honors Undergraduate Research Program (SHURP) at Harvard was initiated in 1991 to encourage minority college students to consider research careers in the biomedical sciences. With support by NIH and NSF, this program has grown from 5 members in 1991 to 23-25 annually in recent years. Ninety-seven percent of the SHURP alumni who have finished college have either graduated from, entered, or are planning to enter Ph.D., M.D./Ph.D., or M.D. programs.

FDSRP. Four Directions Summer Research Program (FDSRP) at Harvard Medical School was initiated in 1994 with the enrollment of 6 students. Now entering its 15th year, FDSRP has brought nearly 150 students to Harvard Medical School. The program includes: an 8-week research project under Harvard Medical School faculty; weekly lecture seminars covering topics such as applying to medical and graduate schools, Native American health care issues, and careers in medicine and biological research, and a final research project presentation to students, directors, and faculty.

SURF. Summer Undergraduate Research Fellowship (SURF) is a 10-week research experience that supports undergraduate students for the summer at Boston University. Participants from institutions across the country are paired with BU faculty members who will serve as their research mentors. As part of the program, participants also attend a series of summer enrichment workshops whose topics range from laboratory safety to research ethics to scientific writing. At the conclusion, SURFers are asked to present a 10-15 minute talk about their research to an audience consisting of their peers, other students, faculty, staff, and invited guests.

CAMP. The CAMP Summer Research Program is an eight-week, full-time research experience. Students work as research assistants on projects that are supervised by a faculty mentor. Student activities include: training in how to write and present a research paper; GRE preparation instruction; field trips to local companies, research institutions, or campus labs; skill building workshops; social events; a required oral presentation at the UCSD Summer Research Conference; and a presentation of their research at the CAMP Statewide Symposium (Winter Quarter).

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

BBRI. As part of its Science Education Outreach Program, Boston Biomedical Research Experience invites Boston-area students into the laboratories for internships each summer. BBRI student interns are given the opportunity to gain hands-on experience using lab equipment and assisting scientists with day-to-day lab activities and to be immersed in real-world, cutting edge research science experiments.

We believe that the participation by the CTCHGV principal investigator and other investigators in the center in existing programs with demonstrated success will strengthen our own program's efforts and the efforts of NHGRI to achieve maximum impact on the professional advancement of underrepresented minorities.

B.4. Postbac Training Program. An established path to success for undergraduate students seeking placement in graduate programs is to transition from an undergraduate research program into a one year postbac training program. The CTCHGV MAP postbac training program is designed to enable trainees to build upon their early research experiences and facilitate placement in a graduate program.

Milestones: We will recruit and enroll one trainee per year in the Postbac Training Program. We will recruit exceptional students with a strong track record in undergraduate research, including trainees from the center's Undergraduate Summer Research Experience Program. Postbac trainees will be encouraged to take GRE prep courses, paid for via program funds, immediately upon their entry into the program and to participate in activities that will strengthen their ability for professional advancement, including presentations, publications and making professional contacts.

C. Management. George Church, the principal investigator of the center, has extensive experience as a mentor for trainees at all professional levels. He assumed oversight of MGIC MAP training program in 2003 and he will continue oversight of the minority training programs for CTCHGV MAP.

Dr. Church has served on the Advisory Committee (on genomics and bioinformatics) for the Society for Advancement of Chicanos and Native Americans in Science (SACNAS) and his lab participated in NCI CURE (Continuing Umbrella of Research Experience), a program designed to increase the number of underrepresented minorities engaged in biomedical cancer research.

While the principal investigator and other key personnel are committed to actively participating in the CTCHGV MAP training program, we will assign personnel specifically devoted to the day-to-day administration of center initiatives.

C.1 Administrative Coordinator of Minority Initiatives. Yveta Masarova is a veteran administrator in the Church lab. She has developed extensive experience in the administration of multi-institution research endeavors and academic environments. She is in continual contact with the principal investigator, investigators and key personnel of the center, program alumni, and administrators on other campuses.

As the Administrative Coordinator of Minority Initiatives, Ms. Masarova will manage day-to-day administration of CTCHGV CEGS MAP activities, including responding by phone, email, and mail communications regarding our program; maintaining and updating center records related to MAP activities and alumni; updating web site information; producing materials for recruitment activities and other program events; and coordinating MAP activities at the institutions of center collaborators. Ms. Masarova also coordinates very closely with the Postdoctoral coordinator in attending Boston area MAP meetings. Ms. Masarova will track center alumni for a minimum of five years following the completion the program via email on an annual basis.

C.2 Postdoctoral Coordinator of Minority Initiatives. As postdoctoral coordinator of MAP initiatives, Dr. Adeyemi Adesokan, a fellow in the Church Lab, is involved with other managers of Boston Area MAP initiatives in planning peer-to-peer outreach events for area students. He also will continue to participate in quarterly teleconference calls and spearhead recruitment efforts aimed at identifying undergraduate summer students and postdoctoral fellows with Yveta Masarova. Finally, he will represent the center at the biannual MAP coordinator meetings, as scheduled by the NIH.

Milestones: At the end of year 1, CTCHGV CEGS will recruit and hire a qualified individual to replace Adeyemi Adesokan as he advances his professional career. The transfer of responsibilities between Dr. Adesokan and his replacement will be done in a manner that ensures overlap and continuity for the program and enables Dr. Adesokan to focus his attention activities necessary for his professional advancement.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

C.3 Ongoing Program Evaluation

In addition to annual reviews of MAP programs, between the second and third years the administrative coordinator of minority initiatives will convene a meeting with the principal investigator and other key staff at the center to evaluate the effectiveness of our programs to enable mid-course corrections, if needed. At this meeting, the following questions will be addressed: (1) Are we meeting our recruitment goals for postdocs, postbacs, and undergraduate students? (2) What percentage of enrolled trainees complete the programs? (3) Have we dedicated sufficient administrative resources to MAP activities? (4) How can we integrate feedback from trainees and mentors to improve our programs? (5) Which investigators are engaged in mentoring trainees, what successes or problems are they having, and can we engage additional investigators in our activities? (6) What has been the impact of collaborating with existing minority programs?

D. MGI CEGS MAP Progress Report

Our evaluation of the MGIC MAP programs identified the postdoc, postbac, and undergraduate training activities (See D.1-D.3 below) as the highest impact of all proposed activities. Therefore, our continuation request proposes to focus on these three activities.

We have modified significantly our "Post College Opportunities" program based on our experiences with MGI CEGS. Initially, we hoped to support 6 postbacs per year, which we have been unable to achieve. Therefore, we have scaled this program down to one student per year. One concern of the principal investigator with the postbac training program is that unless students arrive with exceptional experience from an undergraduate program, a one-year postbac program may be insufficient for placement in graduate school. Furthermore, postbacs typically begin applying to graduate school immediately upon enrollment in the training program. Unless these students already have extensive research experience and a relationship with the PI, letters of recommendation, publications, and other achievements will only be obtained after they have submitted applications to graduate school. One worry is the postbacs will therefore require at least two years of support, and may put off applications to grad school and career advancement. The postbac program will be monitored closely during the first two years, and will focus on enrolling at most one exceptional student per year.

D.1 MGI CEGS MAP Postdoc Program

MGIC MAP is actively supporting one postdoctoral fellow, Adeyemi Adesokan Ph.D, who has been pursuing research under the mentorship of George Church at Harvard Medical School since September 2007. His research in the Church lab is in two directions, basic science and science policy. These include the application of Flux Balance Analysis (FBA) methods to predict conditions necessary for the optimization of ethanol production from cellulose in *Clostridium Phytofermentans* and the development of genomics technologies for pathogen detection and disease surveillance. In collaboration with Prof. Calestous Juma at the Harvard University Kennedy School of Government, he is also investigating policy and capacity building questions related to the impact of genomics and biotechnology in innovation and sustained economic development in the developing world. Adeyemi received a prestigious Private Source

Private Source grant in October 2008 to support his work both in basic science research and science policy. In addition, Yemi has given a number of invited lectures both nationally and internationally about his postdoctoral work. Yemi was named a Private Source Fellow by the Washington DC based Private Source in December 2008.

MGI CEGS MAP Postdoc Success Story Notable success from the MGI CEGS MAP training program include the alumnus Janice A. Lansita, Ph.D. She was a postdoctoral trainee in the James Sherley Lab in 2005 and has since advanced to a research position as a toxicologist at Biogen-IDEA (Cambridge, MA). Dr. Adeyemi Adesokan currently a MAP postdoc in the Church lab was a recipient of a Private Source Private Source grant to develop functional metagenomics technologies to probe drug resistance infectious diseases in the developing world.

D.2 MGI CEGS MAP Postbac Program

The Church Lab is actively supporting two postbac students, Amanda Chilaka and Gerardo Gonzalez-Guardiola.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

MGI CEGS MAP Postbac Success Story Amanda entered the postbac training program in August of 2006 and was accepted to Northeastern University graduate school starting in the Fall 2008. Her training experience has included polony sequencing, proteomics work with absolute quantification and analysis of biofuel production, as well as an introduction into the computational sciences with newly acquired programming skills in Python. She has also been awarded a scholarship through [Private Source] for volunteer work she did to provide medical supplies to hospitals in the Dominican Republic. As a postbac at the Church Lab, Gerardo Gonzalez-Guardiola, has been working with Dr. Francois Vigneault on improving the "miRNA Capture" protocol, with a goal towards the achievement of higher and more specific RNA yields. Gerardo got accepted to the MD/PhD program at the University of Puerto Rico, Medical Science Campus and will be starting next year.

D.3 MGI CEGS MAP Undergraduate Program

Since 2005, MGI CEGS MAP has enrolled 18 summer trainees in the center's Summer Research Experience (see below, Table MAP-2). Three exceptional undergraduate students and their achievements are highlighted below (see below, Success Stories).

Table MAP-2: MGI CEGS MAP Undergraduate Program		
<i>MGIC Trainees enrolled in MGIC Summer Research Experience</i>		
<i>Trainee</i>	<i>Year</i>	<i>MGIC Lab</i>
Personal Info	Summer 2008	Mitra
	Summer 2008	Church
	Summer 2008	Church
	Summer 2008	Church
	Summer 2008	Church
	Summer 2007	Mitra
	Summer 2007	Church
	Summer 2007	Church
	Summer 2007	Church
	Summer 2007	Church
	Summer 2007	Church
	Summer 2006	Mitra
	Summer 2006	Church
	Summer 2006	Church
	Summer 2006	Church
	Summer 2006	Church
	Summer 2005	Mitra
	Summer 2005	Sherley

MGI CEGS MAP Undergraduate Success Stories

Personal Info

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Personal Info

Data and Materials Dissemination Plan

Software, protocol, and data sharing. Following principles also followed in the prior MGI CEGS, CTCHGV will openly share software, protocols and data. Programs and data will be distributed from Church Lab website (<http://arep.med.harvard.edu/>), with data related to human subjects distributed as described below. Programs will be generally available as documented source code with a Harvard and/or UCSD copyright, distributed freely to academics, and with a Harvard and/or UCSD license for commercial use. In line with long-standing Church Lab commitments, we continue to champion concepts that we helped establish for the genome sequencing community that encourage rapid data deposition and technology transfer, such as "Open Source Biology" (OSB) which determines procedures and technologies to aid the distribution of complex reagents more effectively. The related goal of OSB is to prevent exclusive licenses from potentially interfering with technology transfer. In this regard, we will try to move our technology either into the public domain or non-exclusive licensing mechanisms well before they would be normally publishable. Just as genomics has made laudable the publication of loads of "hypothesis neutral" or "negative" results by way of comprehensiveness and data mining/post-experiment hypotheses, so, too, do we hope OSB will encourage a similar process for technology development and transfer. Our current policy for software sharing is closest to ISCB Level 2: "Source code is available for research use to educational institutions, non-profit research institutes, government research laboratories, and individuals, without the right to redistribute". Examples of software that we have shared at Level 2 are available from the Church lab website (see above). In addition to software, the investigators of the proposed center will make many other resources available with a similar level of access. Indeed, the Church Lab is heavily involved with other research institutions in efforts to actively promote ways to make biological resources more 'shareable', such as the synthetic biology Registry of Standard Biological Parts.

CTCHGV human subject data and samples: As described in Research Design (see section 5, Introduction, and section 5.1.2(i)), CTCHGV will use pre-existing, publically available cell lines and tissue samples from HapMap, PGP, Framingham Heart Project, or other sources. Data and derived from these cell lines will be managed and disseminated consistent with NIH policies for restricted access to combinations of phenotype and genomic information that are potentially identifying. To the maximum extent possible, depending on the data source, we will submit information to appropriate publically available data bases such as dbGaP, dbSNP, GEO, the NCBI Trace Archive, etc.

CTCHGV will generate potentially many thousands of altered cell lines from these original samples, including samples for which, for 1000 genes, up to five locations in the cis gene regulatory regions are combinatorially varied within the cells, clonal isolates of these combinatorially altered cells, and similarly altered cell lines that have been generated as or transformed into induced Pluripotent Stem cell (iPS) lines (some differentiated into diverse cell types). While CTCHGV would like to make these thousands of altered cell lines available to the research community, it has not budgeted for their deposition in, maintenance by, and distribution by Coriell or any other repository. CTCHGV will therefore discuss with its Scientific Advisors, with NIH, and with Coriell, the practical prospects and possibilities for maintaining and distributing these altered cell lines. A likely direction will be to deposit an agreed-upon, limited set of key altered cell lines in Coriell, and to seek supplementary funding for maintenance of these altered cell lines and for submission and maintenance of additional sets. The "marked allele" iPS cell lines to be generated in Aim 2 (see Research Design, section 5.2.3) will be high priorities for distribution, as these will comprise an important and potentially unique resource for the research community.

Zinc Finger pools and Zinc Finger Nucleases (ZFN): As described in Aims 1.1 and 4.2 (see Research Design sections 5.1.1 and 5.4.2, respectively for details), CTCHGV plans as part of its work to

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

undertake the generation of a large number of zinc finger pools that can be used to target proteins to specific locations in genomic DNA, and to create thousands of ZFN proteins whose purpose is to induce double stranded breaks and specific genomic DNA locations. CTCHGV plans to disseminate the zinc finger pools and ZFN proteins through the Zinc Finger Consortium, a group of 16 academic laboratories dedicated to making zinc finger technology available to the academic research community. The Zinc Finger Consortium and Dr. Joung both have a strong track record of making all zinc finger engineering methods, software, and reagents broadly available to academic laboratories (see Pearson, *Nature* 2008 (see References) and <http://www.zincfingers.org>). In the case of the ZFNs, CTCHGV will consult and agree with the Zinc Finger Consortium on an appropriate method of distributing materials. As has been done previously for other zinc finger-related resources, all software and protocols will be posted on the Zinc Finger Consortium website (<http://www.zincfingers.org/software-tools.htm> and <http://www.zincfingers.org/protocols.htm>). Plasmids and/or DNA sequence information will be provided through the non-profit plasmid distribution service Addgene, a group with whom the Zinc Finger Consortium has worked with for years in distributing these types of reagents (see: <http://www.addgene.org/zfc>).

Commercialization: As described above, CTCHGV will pursue open and non-exclusive licensing agreements that encourage innovations to be made widely available to researchers and commercial entities. Professor Church has been on the Harvard-wide Copyright and Patent Committee (CPC) for years, a recipient of numerous successful patents, and is in constant contact with the HMS Office of Technology Licensing (OTL). More generally, we will encourage close relationships with companies who can promote broad usage of innovations by incorporating them with other technology into readily usable packages and applications. Professor Church is currently on Scientific Advisory Boards of fifteen companies and has maintained close relationships with many others. He will use these close relationships to encourage companies to adapt CTCHGV innovations into their products, as he did in the prior MGI CEGS.

Current company Scientific Advisory roles of Professor Church

1. *Alacris Pharmaceuticals*, Berlin 2008 (2008-present; Cancer genomics & systems biology)
2. *Danaher-Polonator*, Salem, NH 2007 (2007-present; sequencing system)
3. *Qteros* (formerly *SunEthanol*), Amherst MA 2007 (2007-present; Biofuels)
4. *LightSpeed Genomics*, (2007-present; high-speed DNA sequencing)
5. *Complete Genomics*, Sunnyvale, CA 2006 (2006-present; sequencing nanoarrays)
6. *Knome, Inc.*, Cambridge, MA (2007-present; Human Genome Sequencing)
7. *LS9*, San Francisco, CA 2005 (2006-present; Biologically engineered fuels)
8. *Enzymatics*, Beverly, MA 2006 (2006-present; Large-scale, high quality enzymes)
9. *IntelligentBioSystems (IBS)*, Waltham, MA 2006 (2006-present; Sequencing by Extension on beads)
10. *PharmoRx*, Wellesley, MA 2005 (2005-present; secure medication)
11. *Pacific Biosciences*, Menlo Park, CA 2004 (2008-present; real-time single-molecule sequencing)
12. *Helicos Biosciences, Corp.*, Cambridge, MA 2004 (2003-present; Single-molecule DNA sequencing)
13. *23andme*, Mountain View, CA 2006 (2006-present; personal genomics)
14. *DNAdirect*, San Francisco, CA 2004 (2006-present; DNA diagnostics)
15. *Genomatica*, San Diego CA 2001 (2001-present; microbial metabolic models)

Past Companies licensing Church lab patents or software

1. *Affymetrix (Affymax)*, Palo Alto, CA 1993 (1990-present; Oligonucleotide arrays)
2. *Agencourt* (now *Beckman Coulter*), Beverly, MA 2000 (2003-2006; Polony bead sequencing by ligation)
3. *Applied Biosystems*, Foster City, CA 1981 (2003-2006; via Agencourt, polony bead sequencing by ligation)
4. *Lynx - Solexa Illumina*, Hayward, CA 1992 (2000-2006; multiplex tags)
5. *Pyrosequencing*, (also *Biotage - 454*), Stockholm 2000 (2001-present; modified dNTPs for array sequencing)
6. *Millipore*, Bedford, MA (1989-1990; multiplex sequencing)

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

7. *Mosaic Technologies*, Boston, MA 1994 (1993-94 & 2000-2001; DNA diagnostics)
8. *Agilent*, 2000 (2001-present; nucleic acid nanopore sensors)

It should be noted that Professor Church has given up financial incentives from all sequencing and personal genomics companies (see Management and Organization Plan, G. Conflict of Interest Policy).

Inclusion Enrollment Report

Not applicable.

Protection of Human Subjects

Evaluative Info

Our proposed research does not involve a Clinical Trial.

Our proposed research also does not involve an NIH-Defined Phase III Clinical Trial.

In accord with regulations regarding all research that falls under Human Subjects Research Exemption 4, NIH policies for inclusion of women, minorities, and children in clinical research, and targeted/planned enrollment tables, do not apply to this proposal, and therefore we are not including the corresponding sections as part of our submission.

Inclusion of Women and Minorities

Not applicable. (See Protection of Human Subjects.)

Targeted/Planned Enrollment Table

Not applicable. (See Protection of Human Subjects.)

Inclusion of Children

Not applicable. (See Protection of Human Subjects.)

Vertebrate Animals

Not applicable. The research proposed does not involve vertebrate animals.

Select Agents

Not applicable. The research proposed does not involve Select Agents.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Multi-PI/PD Leadership Plan

Not applicable.

Resource Sharing Plan

Our proposed CTCHGV CEGS will generate several resources of potentially great value to the research community which we plan to make available to the community to the extent possible. We have described these plans and resources in our Data and Materials Dissemination Plan (DMDP) and briefly summarize them here.

Open-source policy: The Church Lab adheres to open-source standards to the extent possible as a matter of policy, for software, data, materials, and technology generally. Please refer to the DMDP for additional details.

CTCHGV software and data resources: Software developed by CTCHGV will be made available from the Church Lab web site in accordance with its open source policy. Data derived from human samples will be placed on appropriate public repositories to the extent possible in a manner consistent with NIH policies for restricted access to information that is potentially identifying. Please refer to the DMDP for additional details.

CTCHGV human cell lines: CTCHGV will obtain pre-existing, publically available human samples from a small numbers of subjects but will potentially generate many thousands of cell lines with specific alterations engineered from these original samples, including induced Pluripotent Stem cell (iPS) lines (some differentiated into diverse cell types). CTCHGV will consult with its Scientific Advisors, with NIH, and with Coriell to determine a strategy for distribution of these altered cell lines. A likely direction will be to deposit an agreed-upon, limited set of key altered cell lines in Coriell, and to seek supplementary funding for maintenance of these altered cell lines and for submission and maintenance of additional sets. A high priority for distribution will be the "marked allele" iPS cell lines to be derived in Aim 2 (Research Design, section 5.2.3), which will comprise an important and potentially unique resource for the research community. Please refer to the DMDP for additional details.

Zinc Finger pools and Zinc Finger Nucleases (ZFN): As described in Aims 1.1 and 4.2 (see Research Design sections 5.1.1 and 5.4.2, respectively for details), CTCHGV plans as part of its work to undertake the generation of a large number of zinc finger pools that can be used to target proteins to specific locations in genomic DNA, and to create thousands of ZFN proteins whose purpose is to induce double stranded breaks and specific genomic DNA locations. CTCHGV plans to disseminate the zinc finger pools and ZFN proteins through the Zinc Finger Consortium, a group of 16 academic laboratories dedicated to making zinc finger technology available to the academic research community. The Zinc Finger Consortium and Dr. Joung both have a strong track record of making all zinc finger engineering methods, software, and reagents broadly available to academic laboratories (see Pearson, *Nature* 2008 (see References) and <http://www.zincfingers.org>). In the case of the ZFNs, CTCHGV will consult and agree with the Zinc Finger Consortium on an appropriate method of distributing materials. As has been done previously for other zinc finger-related resources, all software and protocols will be posted on the Zinc Finger Consortium website (<http://www.zincfingers.org/software-tools.htm> and <http://www.zincfingers.org/protocols.htm>). Plasmids and/or DNA sequence information will be provided through the non-profit plasmid distribution service Addgene, a group with whom the Zinc Finger Consortium has worked with for years in distributing these types of reagents (see: <http://www.addgene.org/zfc>).

Commercialization: To broaden the availability to the research community of innovations developed by CTCHGV, the Church Lab will work with the Harvard Medical School Office of Technology Licensing to obtain open and non-exclusive licenses that will encourage commercialization of these innovations. Please refer to the DMDP for additional details.

Consortium/Contractual Agreements

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

The applicant organization will be Harvard Medical School (HMS) as the director and a significant fraction of the center is located there. The contractual investigators and their institutions are in agreement with this arrangement. The actual sub-contracts for individual institutions and investigators are provided elsewhere in this application and briefly described here.

- Dr. George Q. Daley, Children's Hospital, will work with the Church Lab on development of methods for generating, engineering, maintaining, differentiating, and analyzing human induced Pluripotent Stem cells for Aim 2 (see Research Design, section 5.2).
- Dr. Robert Green, Boston University, will be a consultant for the Center for training (see Training and Minority Action Plans) both Center and non-Center researchers on potential uses of Center methods for understanding human disease associations, and on human subjects research issues.
- Dr. J. Keith Joung, Massachusetts General Hospital, will work with the Church Lab on developing and optimizing a very large repository of zinc finger nucleases (ZFNs) for use in engineering human cells for Aims 1.1 and 4.2 (see Research Design, sections 5.1.1 and 5.4.2).
- Professor Kun Zhang, University of California at San Diego, will work with the Church Lab on determination of the causality of cis variations for 1000 genes in Aim 1 (see Research Design, section 5.1).

DAVID ALTSHULER, M.D., PH.D.

Director, Program in Medical and Population Genetics
Broad Institute of Harvard and MIT

Professor of Genetics and Medicine
Harvard Medical School
Massachusetts General Hospital

Thursday, May 7, 2009

George Church, PhD
Professor of Genetics
Harvard Medical School

Re: CEGS Center for Transcriptional Consequences of Human Genetic Variation

Dear George:

I am truly excited to be an active collaborator and Scientific Advisor to your proposed CEGS Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV).

In my opinion, two of the most significant developments from the Human Genome Project are (a) the recognition of that most of the conserved DNA in the human genome is non-coding (and of unknown function), and (b) that systematic studies of human genome variation can be used to connect specific genomic segments with human disease. Unfortunately, while these two developments have raced ahead, biological progress has been stymied by our inability to read and predict the functional consequences of non-coding DNA variation.

Your CEGS aims to tackle this problem in an audacious and innovative manner. The four aims are each exciting, and bring in another of the most exciting developments in biomedical science: iP_s technology. Your vision is truly breathtaking, as the previous cycle of your CEGS and the proposed will move us towards a world in which a complete suite of human tools for investigation will be available: genetic screens in humans based on correlation of phenotype with complete genotype; knowledge of the "regulatory code" by which we can interpret non-coding variation, and the ability to study the effect of specific genotypes *in vitro* based on controlled differentiation of iP_s cells.

It will be my great pleasure to provide my own active engagement and expertise in any way that your group would find helpful. As you know, my own research involves studies of human genetic variation, and its use to perform genetic mapping of disease genes in man. My lab has identified dozens of novel loci for type 2 diabetes, lipids, myocardial infarction, and other diseases, and we are actively performing fine mapping and deep sequencing studies of these loci. I am Co-Chair of the Steering Committee of the 1000 Genomes Project, and this effort is providing a richer catalogue of genetic variation that will soon be associated with phenotypes. My own lab is deeply interested in the Aims of your CEGS, and we can't wait to get started on collaborations with the Center!

Sincerely,
Signature

David M. Altshuler, M.D., Ph.D.



Phone: 617-726-5940 Broad Institute of Harvard and MIT
Fax: 617-726-5937 7 Cambridge Center
E: altshuler@molbio.mgh.harvard.edu Cambridge, MA 02142



Uday Gupta
President & CEO
Global Cell Solutions, Inc.
770 Harris Street, Suite 104
Charlottesville, VA 22903

January 12, 2009

770 Harris Street
Suite 104
Charlottesville, VA 22903
T 434-975-4271
F 734-661-4707
uday.gupta@globalcellsolutions.com
www.globalcellsolutions.com

Church Laboratory
Harvard Medical School
Department of Genetics
NRB Rm 238
77 Ave. Louis Pasteur
Boston, MA 02115

Dear Professor Church:

We are writing to express our enthusiastic commitment to collaborating with you on a scalable and automated microcarrier-based cell culture expansion and cryopreservation system for iPS cell culture generation and maintenance in connection with your CEGS proposal for a "Center for Transcriptional Consequences of Human Genetic Variation."

As discussed, we will provide expertise and materials for your work on development and use of this system in your CEGS, while you attempt to generate and maintain thousands of engineered iPS cell lines.

We believe that your testing and adaptation of this system will help us develop and demonstrate it as an effective method for managing stem cell lines that will have wide applicability and will greatly benefit biomedical research.

Warmest Regards,

Signature



Uday Gupta
President



**HARVARD
MEDICAL SCHOOL**



**BRIGHAM AND
WOMEN'S HOSPITAL**

77 Avenue Louis Pasteur, Suite 168 • Boston, MA 02115 • tel: 617 525-4451 • fax: 617 525-4488 • rplenge@partners.org

Robert M. Plenge, M.D., Ph.D.
Assistant Professor of Medicine
Harvard Medical School

Director, Genetics & Genomics
Division of Rheumatology, Immunology and Allergy
Brigham and Women's Hospital

May 14th, 2009

George Church, Ph.D.
77 Ave. Louis Pasteur
Harvard Medical School
Boston, MA 02115 USA

Dear George,

I am pleased to collaborate with you in your proposed CEGS Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV). I am interested in continuing our collaborations applying your methods to see if we can shed light on my research on sequencing-based association studies in common autoimmune diseases such as rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE). I'd be happy to advise and provide assistance in relating your methods to GWAS. I will also be happy to be a Scientific Advisor to your Center.

Best regards,

Signature

Robert Plenge, M.D., Ph.D.



Howard Hughes Medical Institute
Research Laboratories

Stephen J. Elledge, Ph.D.
Investigator

May 14, 2009

Dear George,

I am very pleased to collaborate with you and Keith Joung in your proposed CEGS Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV). I will be happy to provide you access to expertise and reagents for genome wide RNAi and overexpression screens in human cells, and believe that these can contribute greatly to your success in optimizing homologous recombination in human cells. As you know I have already been working in this direction and believe that your use of this work will be an excellent application of it. I will also be very happy to serve as a Scientific Advisor to your Center.

Sincerely,

Signature

Stephen J. Elledge, Ph.D.

Harvard Medical School, Department of Genetics
77 Avenue Louis Pasteur, NRB Room #158D, Boston Massachusetts 02115
(617) 525-4510 • Fax (617) 525-4500
selledge@genetics.med.harvard.edu

**Harvard Medical School
Department of Genetics**

77 Avenue Louis Pasteur
Boston, MA, USA 02115
<http://arep.med.harvard.edu>



May 15, 2009

Dear George,

I am pleased to collaborate with you in your proposed CEGS Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV).

As you know, the ability to search the human genome for common polymorphisms that associate with disease risk has overwhelmingly implicated the non-protein-coding parts of the human genome as harboring phenotypically influential genetic variation in humans. Based on both published genome-wide association studies and ongoing resequencing of the genes near association signals, it already appears likely that a majority of these influences of human genetic variation on disease risk will not be explained by variation in protein-coding sequence. Our limited ability to interpret the functional significance of variation in regulatory sequence currently limits our understanding of the biological meaning of these results and the molecular etiology of disease.

The methods you are developing will enable a much deeper understanding of how regulatory variations lead to phenotypes at many levels – from gene expression in cells and tissues to disease risk in populations. This fits extremely well with my interests in genetically complex disease, regulatory polymorphism and genome structural variation.

I will be happy to provide expertise and assistance in understanding the relationships among these phenomena. I will also be interested in seeing the extent to which your methods can help reveal the functional significance of structural polymorphism in non-coding regions of the human genome.

Sincerely,

Signature

Steven A. McCarroll, Ph.D.
Assistant Professor of Genetics

Complete Genomics

May 14, 2009

George Church, Ph.D.
Department of Genetics
Harvard Medical School
New Research Building, Room 238
77 Avenue Louis Pasteur
Boston, MA 02115

Dear Dr. Church:

We are pleased to collaborate with you in your proposed CEGS Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV). In support of your work, we will be happy to provide diploid genome sequences of 10 human subjects from HapMap, Personal Genome Project, Framingham, or other sources. We believe your use of these sequences to identify gene regulatory region variations that control transcription will provide excellent demonstrations of the utility and high potential of diploid genome sequences. We will also be very interested in collaborating with you on your development of single cell transcriptomics and integration of such data with the CGI genomes.

Complete Genomics has already demonstrated the ability to accurately and efficiently sequence the entire human genome (<http://www.completegenomics.com/dataRelease/>). CGI offers sequencing services to provide $\geq 90\%$ completeness and 120GB of 35bp+35bp paired-end reads on normal (i.e. non-tumor) human DNA. Because of the scale involved, CGI provides access to this 3rd generation sequencing technology through services performed at our Mountain View Facility. Our technology is more thoroughly described in the "Complete Genomics Technology Paper" found here:

<http://www.completegenomics.com/pages/materials/CompleteGenomicsTechnologyPaper.pdf>

Deliverables:

- a) Sample shipment: 100 μ g of high-molecular weight genomic DNA at a concentration of 100 – 200ng/ml aliquoted into bar-coded 96-well plates supplied by Complete Genomics. All samples should be shipped on dry ice.
 - Quantification by the Pico-green assay (preferably with the Quant-iTTM PicoGreen[®] ds DNA kit from Invitrogen)
 - Complete Genomics will accept deliveries Monday through Friday between 8:00am and 5:00pm Pacific Time (excluding U.S. public holidays). To ensure timely delivery, we recommend that samples be shipped Monday through Wednesday.
- b) Sequencing: Complete Genomics will perform library construction, sequencing, data mapping and data assembly in accordance with its standard processes. For normal (i.e. non-tumor) human DNA, Complete Genomics will provide ≥ 120 GB of 35bp+35bp paired-end reads, for an average of 40X coverage and 90% completeness across the genome.

Page 1 of 2

2071 Stierlin Court • Mountain View, CA 94043 • Tel: (650) 943-2800 • Fax: (650) 964-2108



- c) Data Delivery – (See Data File Formats document for more detail)
 - 35bp+35bp paired-end reads file
 - Variation File
 - Data delivered on one or more encrypted hard disk drives. Encryption key(s) to be provided under separate cover.

We look forward to supporting your project. Please feel free to contact me with any further questions.

Best regards,

Signature

A rectangular box with a thin black border, intended for a signature. The word "Signature" is printed in the top-left corner of the box.

Aaron Solomon
Vice President, Business Development

Main - Harvard

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

CHECKLIST**TYPE OF APPLICATION** (Check all that apply.)

- ☒ NEW application. (This application is being submitted to the PHS for the first time.)
- ☐ RESUBMISSION of application number: _____
(This application replaces a prior unfunded version of a new, renewal, or revision application.)
- ☐ RENEWAL of grant number: _____
(This application is to extend a funded grant beyond its current project period.)
- ☐ REVISION to grant number: _____
(This application is for additional funds to supplement a currently funded grant.)
- ☐ CHANGE of program director/principal investigator.
Name of former program director/principal investigator: _____
- ☐ CHANGE of Grantee Institution. Name of former institution: _____
- ☐ FOREIGN application ☐ Domestic Grant with foreign involvement List Country(ies) Involved: _____

INVENTIONS AND PATENTS (Renewal appl. only) ☐ No ☒ YesIf "Yes," ☒ Previously reported ☐ Not previously reported**1. PROGRAM INCOME** (See instructions.)

All applications must indicate whether program income is anticipated during the period(s) for which grant support is request. If program income is anticipated, use the format below to reflect the amount and source(s).

Budget Period	Anticipated Amount	Source(s)

2. ASSURANCES/CERTIFICATIONS (See instructions.)

In signing the application Face Page, the authorized organizational representative agrees to comply with the policies, assurances and/or certifications listed in the application instructions when applicable. Descriptions of individual assurances/certifications are provided in Part III and listed in Part I, 4.1 under Item 14. If unable to certify compliance, where applicable, provide an explanation and place it after this page.

3. FACILITIES AND ADMINISTRATIVE COSTS (F&A)/ INDIRECT COSTS. See specific instructions.

- ☒ DHHS Agreement dated: 2/27/08 ☐ No Facilities And Administrative Costs Requested.
- ☐ DHHS Agreement being negotiated with _____ Regional Office.
- ☐ No DHHS Agreement, but rate established with _____ Date _____

CALCULATION* (The entire grant application, including the Checklist, will be reproduced and provided to peer reviewers as confidential information.)

a. Initial budget period:	Amount of base \$	<u>2,008,582</u>	x Rate applied	<u>69.5</u>	% = F&A costs	\$	<u>1,395,964</u>
b. 02 year	Amount of base \$	<u>1,938,314</u>	x Rate applied	<u>69.5</u>	% = F&A costs	\$	<u>1,347,128</u>
c. 03 year	Amount of base \$	<u>1,936,874</u>	x Rate applied	<u>69.5</u>	% = F&A costs	\$	<u>1,346,128</u>
d. 04 year	Amount of base \$	<u>1,935,456</u>	x Rate applied	<u>69.5</u>	% = F&A costs	\$	<u>1,345,142</u>
e. 05 year	Amount of base \$	<u>1,933,995</u>	x Rate applied	<u>69.5</u>	% = F&A costs	\$	<u>1,344,126</u>
TOTAL F&A Costs						\$	6,778,488

*Check appropriate box(es):

- ☐ Salary and wages base ☒ Modified total direct cost base ☐ Other base (Explain)
- ☐ Off-site, other special rate, or more than one rate involved (Explain)

Explanation (Attach separate sheet, if necessary.):

Indirect cost:

Harvard : 3 Subs: (3x25,000=75,000)

MAP Indirect: \$1,101,548

MAIN Indirect: \$6,778,488

4. DISCLOSURE PERMISSION STATEMENT: If this application does not result in an award, is the Government permitted to disclose the title of your proposed project, and the name, address, telephone number and e-mail address of the official signing for the applicant organization, to organizations that may be interested in contacting you for further information (e.g., possible collaborations, investment)? ☐ Yes ☒ No

MAR

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

CHECKLIST**TYPE OF APPLICATION** (Check all that apply.)☒ NEW application. (This application is being submitted to the PHS for the first time.)☐ RESUBMISSION of application number: _____

(This application replaces a prior unfunded version of a new, renewal, or revision application.)

☐ RENEWAL of grant number: _____

(This application is to extend a funded grant beyond its current project period.)

☐ REVISION to grant number: _____

(This application is for additional funds to supplement a currently funded grant.)

☐ CHANGE of program director/principal investigator.

Name of former program director/principal investigator: _____

☐ CHANGE of Grantee Institution. Name of former institution: _____☐ FOREIGN application ☐ Domestic Grant with foreign involvement List Country(ies) Involved: _____INVENTIONS AND PATENTS (Renewal appl. only) ☐ No ☒ YesIf "Yes," ☒ Previously reported ☐ Not previously reported**1. PROGRAM INCOME** (See instructions.)

All applications must indicate whether program income is anticipated during the period(s) for which grant support is request. If program income is anticipated, use the format below to reflect the amount and source(s).

Budget Period	Anticipated Amount	Source(s)

2. ASSURANCES/CERTIFICATIONS (See instructions.)

In signing the application Face Page, the authorized organizational representative agrees to comply with the policies, assurances and/or certifications listed in the application instructions when applicable. Descriptions of individual assurances/certifications are provided in Part III and listed in Part I, 4.1 under Item 14. If unable to certify compliance, where applicable, provide an explanation and place it after this page.

3. FACILITIES AND ADMINISTRATIVE COSTS (F&A)/ INDIRECT COSTS. See specific instructions.☒ DHHS Agreement dated: 02/27/08☐ No Facilities And Administrative Costs Requested.☐ DHHS Agreement being negotiated with _____ Regional Office.☐ No DHHS Agreement, but rate established with _____ Date _____

CALCULATION* (The entire grant application, including the Checklist, will be reproduced and provided to peer reviewers as confidential information.)

a. Initial budget period:	Amount of base \$	300,000	x Rate applied	69.5	% = F&A costs	\$	208,250
b. 02 year	Amount of base \$	308,280	x Rate applied	69.5	% = F&A costs	\$	214,255
c. 03 year	Amount of base \$	316,808	x Rate applied	69.5	% = F&A costs	\$	220,182
d. 04 year	Amount of base \$	325,593	x Rate applied	69.5	% = F&A costs	\$	226,287
e. 05 year	Amount of base \$	334,640	x Rate applied	69.5	% = F&A costs	\$	232,575
TOTAL F&A Costs						\$	1,101,548

*Check appropriate box(es):

☐ Salary and wages base☒ Modified total direct cost base☐ Other base (Explain)☐ Off-site, other special rate, or more than one rate involved (Explain)

Explanation (Attach separate sheet, if necessary.):

Indirect cost:**4. DISCLOSURE PERMISSION STATEMENT:** If this application does not result in an award, is the Government permitted to disclose the title of your proposed project, and the name, address, telephone number and e-mail address of the official signing for the applicant organization, to organizations that may be interested in contacting you for further information (e.g., possible collaborations, investment)? ☐ Yes ☒ No

From: John Aach
To: Schloss, Jeff (NIH/NHGRI) [E]: "George Church"
Cc: Chen, Debbie P (NIH/NHGRI) [E]: "Yveta Masarova"
Subject: RE: items for CEGS award
Date: Wednesday, June 16, 2010 9:28:05 AM

Hi Jeff,

We have worked through (1) Other Support, and have just completed reviewing and updating (3) Equipment, and you should have our responses soon if you do not have them already.

The remaining item is (2) Stem Cells, and your understanding is precisely correct: We are not planning to use human embryonic stem cells, but we are planning to use human iPS as well as a selection of publically available human-derived cell lines such as those you indicate (HapMap, PGP, Framingham HS,...).

Thanks very much,
--- John Aach

From: Schloss, Jeff (NIH/NHGRI) [E] [mailto:schlossj@exchange.nih.gov]
Sent: Monday, May 31, 2010 4:44 PM
To: George Church (gmc@harvard.edu)
Cc: 'John Aach'; Chen, Debbie P (NIH/NHGRI) [E]
Subject: items for CEGS award

George,

Now that the MAP issues are settled we are working on the new CEGS award (1 P50 HG005550-01). I have the following requests now. As Debbie reviews the budget she may have additional questions.

1. Other support: Please submit Just-in-time Other Support information for the four Key Personnel.

2. Stem Cells: Are you planning to use any human embryonic stem cells in the course of this work?

I understand that you will be creating human iPS cells. These don't fall under current NIH stem cell regulations. You will also be using publicly-available human-derived cell lines (e.g., HapMap, PGP, Framingham HS, etc.).

3. Equipment: Please update your equipment request list and, provide additional justification per the review panel, taking into account how your plans may have changed since submitting the application, and equipment that may already be available in your lab to fill some of these needs.

We will restrict the equipment funds so that you may purchase any of the items on the list you will provide (pending our review/approval of that list prior to making the award). If you decide to use the funds for something else that's not on the list you'll need to check with us first.

We anticipate that you will need to expend much of the equipment funds during the first year of the grant, but that you may decide to carry some of the funds forward to meet unanticipated needs in the future years.

Please let me know if you have any questions,

Best regards,
Jeff

Jeffery A. Schloss, Ph.D.
Program Director
Technology Development Coordination
National Human Genome Research Institute
National Institutes of Health
Suite 4076, MSC 9305
5635 Fishers Lane
Bethesda, MD 20892-9305
COURIER SERVICES SHOULD USE:
Rockville, Maryland 20852
tel: 301-496-7531
fax: 301-480-2770
e-mail: jeff_schloss@nih.gov

Writing or renewing an NIH grant application? New SF424 R&R forms require specific versions of Adobe Reader to ensure the forms work properly. Incorrect versions may prevent successful submission of your application to Grants.gov. See NIH's [eSubmission Web site](#) for more information.

From: [Chick, Cheryl \(NIH/NHGRI\) \[E\]](#)
To: [Oken, Lisa \(NIH/NHGRI\) \[E\]](#)
Subject: FW: church_P50HG005550
Date: Monday, October 25, 2010 12:05:09 PM
Attachments: Church update P50HG005550.pdf

FYI – I'm copying you on a response/intro that you are the new GM....

From: Chien, Kathleen [mailto:Kathleen_Chien@hms.harvard.edu]
Sent: Monday, October 25, 2010 12:01 PM
To: Schloss, Jeff (NIH/NHGRI) [E]; Schloss, Jeff (NIH/NHGRI) [E]
Cc: Chen, Debbie P (NIH/NHGRI) [E]; Chick, Cheryl (NIH/NHGRI) [E]; 'George Church'; 'Yveta Masarova'; Williams, Dirk R (forwarding)
Subject: church_P50HG005550

Dear Dr. Schloss,

Attached is the documentation that was requested in the NGA for our newly awarded P50HG005550 grant under the direction of Dr. George Church.

Specially included:

1. IRB approval letter from HMS.
2. IRB approval letter from UCSD.
3. Certification of Human Subjects education for key personnel: Dr. Church, Dr. Zhang, Dr. Daley, Dr. Joung.
4. Updated face page for HMS.
5. Updated face page for UCSD.
6. Updated justification for Consultants.

Dr. Dae Kim has left HMS and will not be part of the project.

Please let me know if you have any questions or need additional information.

Thank you.

Kathleen Chien
Senior Sponsored Research Administrator
Sponsored Programs Administration
Harvard Medical School
Gordon Hall, Suite 509
25 Shattuck St.
Boston, MA 02115
617-432-5801
fax: 617-432-2651
kathleen_chien@hms.harvard.edu

HARVARD MEDICAL SCHOOL
OFFICE FOR RESEARCH SUBJECT PROTECTION



COMMITTEE ON HUMAN STUDIES
188 LONGWOOD AVENUE, SUITE 111
BOSTON, MASSACHUSETTS 02115
TEL: (617) 432-3071
FAX: (617) 432-5175

George Church, PhD
Harvard Medical School
Genetics NRB Room 238
77 Avenue Louis Pasteur
Boston, MA 02115
Email: gmc@harvard.edu

RE: Causal Transcriptional Consequences of Human Genetics Variation (CTCHGV)
CHS Study Number: M18770-101
Review Date: April 27, 2010

Dear Dr. Church,

This letter certifies that the INITIAL materials for the grant identified above have been reviewed by the Faculty of Medicine Committee on Human Studies. This Committee has given due consideration to each of the items on page 2 and attached hereto, and specifically; (i) the adequacy of protection of the rights and welfare of the subjects involved; (ii) the risks and potential benefits to the subject or importance of the knowledge to be gained; and (iii) the adequacy and appropriateness of the methods used to secure informed consent. Please note that each project under this grant and involving human subjects will require additional and separate review.

Status:	Approved: Full Committee
Effective Approval Date:	April 27, 2010
Study Expiration Date:	April 26, 2011

Funding Source:	National Institutes of Health; P50HG005550-01 <i>Causal Transcriptional Consequences of Human Genetics Variation (CTCHGV)</i>
Approved Sample Size:	N/A
Waivers:	None
Approved Documents:	None

Initial and Continuing approvals are granted for a period of **one year only** and must be renewed annually. Failure to obtain proposal renewal by the expiration date will result in immediate termination of your study with resultant notification of your funding source. **Any modifications** made to this study must be reviewed and approved by the Committee in advance of use. In addition, **adverse events** of any kind must be reported immediately in writing to the Committee, as they occur. Failure to report adverse events or changes to the protocol will also result in immediate termination of your study approval with resultant notification of your funding source.

Please contact your Research Officer, Lucas Breen (lucas.breen@hms.harvard.edu) if you have any questions. Any correspondence with the CHS office regarding this action should note the CHS Study Number indicated at the top of this letter.

Sincerely,
Pamela Richmond
Associate Director, IRB Operations

*This institution has an approved Assurance of Compliance on file with the Department of Health and Human Services.
The Federalwide Assurance number is FWA00007071.*

VIZ00099703

Responsibilities of the Investigator:

1. The Investigator shall obtain the prior approval of the HMS/HSDM Committee on Human Studies (CHS) before implementing *any* protocol changes, *unless* a change is necessary to eliminate apparent immediate hazards to the research participants. If, because of the aforementioned, prior approval was not able to be obtained, the CHS must be informed, in writing, and with specificity, of the deviation from the approved protocol at the first available opportunity and, in no event, later than two (2) business days after discovery. **Failure to obtain prior approval of a change in the protocol will jeopardize your ability to use the data collected during the lapsed period and may result in further disciplinary action.**
2. The Investigator is solely responsible for seeking and obtaining re-approval of her/his research before the expiration date listed on the previous page if s/he wishes to continue the research beyond the stated approval period. Failure to do so will result in a lapse in approval. **If your approval lapses you may not gather any additional data in connection with your research and your sponsor may need to be notified.**
3. Unless an Investigator has specifically asked for, and been granted, a waiver from the requirements of informed consent and/or the documentation of same, an exact duplicate of the **CHS approved and stamped consent form, attached hereto (if applicable), shall be the only form used to enroll participants.** At no time may a participant be enrolled with an expired consent form or one that has been revised, without prior approval of the CHS. In addition, **if an amendment to the consent form has been approved, the newly amended consent form is the *only* form that may be used, going forward, to enroll new participants.**
4. Investigators must submit to the CHS adverse events, unanticipated problems with the study/study participants, non-compliance with the approved protocol, suspensions or terminations of research, complaints about the research (from study participants or others), protocol deviations and violations, within 48 hours, or as soon as learning of the event. Data safety monitoring reports (as applicable) must be submitted when obtained by the Investigator. These event reports must be sent sponsors and appropriate federal agencies as applicable and as required by those entities.

In conducting this review, the CHS Chair or the Chair's designee has given due consideration to ensuring the following:

- Risks to participants are minimized: (i) by using procedures which are consistent with sound research design, and which do not unnecessarily expose participants to risk, and (ii) whenever appropriate, by using procedures already being performed on the participants for diagnostic or treatment purposes.
- Risks to participants are reasonable in relation to anticipated benefits, if any, to participants, and the importance of the knowledge that may reasonably be expected to result from the research.
- Selection of participants is equitable. In making this assessment, the CHS takes into account the purposes of the research and the setting in which the research will be conducted, and is cognizant of the special problems of research involving vulnerable populations, such as children, prisoners, pregnant women, persons with cognitive impairments, or with economic or educational disadvantages.
- Informed consent will be sought from each prospective participant or the participant's legally authorized representative, in accordance with, and to the extent required by §46.116.
- Informed consent will be appropriately documented, in accordance with, and to the extent required by §46.117.
- When appropriate, the research plan makes adequate provision for monitoring the data collected, to ensure the safety of participants.
- When appropriate, there are adequate provisions to protect the privacy of participants, and to maintain the confidentiality of data.
- When some or all of the participants are likely to be vulnerable to coercion or undue influence, such as children, prisoners, pregnant women, persons with cognitive impairments or with economic or educational disadvantages, additional safeguards have been included in the study to protect the rights and welfare of these participants.

For a complete list of responsibilities, policies and procedures, see the CHS website:
<http://www.hms.harvard.edu/orsp/human/human.html>

071575



UNIVERSITY OF CALIFORNIA, SAN DIEGO
HUMAN RESEARCH PROTECTIONS PROGRAM

TO: Kun Zhang Mailcode: 0412

RE: Project #071575
Cis-regulatory DNA polymorphisms in the human genome.

Dear Dr. Zhang:

The above-referenced project was reviewed and approved by one of this institution's Institutional Review Boards in accordance with the requirements of the Code of Federal Regulations on the Protection of Human Subjects (45 CFR 46 and 21 CFR 50 and 56), including its relevant Subparts. This approval, based on the degree of risk, is for 365 days from the date of **IRB review and approval** unless otherwise stated in this letter. The regulations require that continuing review be conducted on or before the 1-year anniversary date of the IRB approval, even though the research activity may not begin until some time after the IRB has given approval.

Approved consent form documents will not be utilized in this study, as it involves the use of samples obtained from a repository and Harvard University.

The IRB reviewed this study and determined that the waiver of informed consent may be granted for this project as it meets the following requirements as outlined in 45 CFR 46.116(d). The research is minimal risk; the waiver or alteration will not adversely affect the rights and welfare of the subjects; and the research could not practicably be carried out without the waiver or alteration. Whenever appropriate, the subjects will be provided with additional pertinent information after participation.

Date of IRB review and approval: 8/19/2010

On behalf of the Institutional Review Board,

Signature

A rectangular box intended for a signature, currently empty.

mb

Michael Caligiuri, Ph.D.
Director, Human Research Protections Program
(858) 455-5050

Note: All Human Subject research conducted at the VA facility and/or utilizing VA/VMRF funds **MUST BE APPROVED** by the VA Research and Development Committee prior to commencing any research. In addition, please ensure that the clinical trial agreement or other funding is appropriately in place prior to conducting any research activities.

VIZ00099705

IRB approval does not constitute funding **or other institutional required approvals**. Should your studies involve other review committees such as Conflict of Interest (COI), Protocol Review Monitoring Committee (PRMC), and committees under Environmental Health & Safety (EH&S) such as Institutional Biosafety Committee (IBC), Human Exposure Committee (HERC), and RSSC (Radiation Safety and Surveillance Committee), it is the researchers responsibility to ensure that all approvals are in place prior to conducting research involving human subjects or their related specimens.

Approval release date: 8/20/2010

UCSD HUMAN RESEARCH PROTECTIONS PROGRAM

GENERAL APPROVAL INFORMATION

The information below does not encompass all human subjects protections requirements, however, is intended to highlight those of significance to ensure awareness by researchers engaged in research involving human subjects or their related specimens and data.

Approval Letters and Consent Documents

Unless otherwise stated, approval letters will be accompanied by stamped, approved consents. Should a study be closed to accrual and no consent released as a result, this information will be documented on the approval letter. Also, any waivers will be documented in the approval letter (such as waiver of documented consent or waiver of authorization for use of PII).

All Human Subject research conducted at the VA facility and/or utilizing VA/VMRF funds MUST BE APPROVED by the VA Research and Development Committee prior to commencing any research. In addition, the PI must ensure approval is in place from other appropriate review boards (such as Radiation Safety, Institutional Biosafety Committee, Conflict of Interest, ESCRO, etc.)

If other institutions are involved, the PI must ensure that IRB approvals (or other administrative approvals) from those sites are secured and forwarded for the study file. In addition, PI's must ensure that the clinical trial agreement, as applicable, or other funding (such as a grant) is appropriately in place prior to conducting any research activities. IRB approval does not constitute funding approval.

Duration of IRB approval

The IRB may grant approval up to 365 days. (See 45 CFR 46.109(d) (DHHS) and 21 CFR 56.109(d) (FDA)). However, for some studies the committee may grant approval for a lesser period or a specific number of subjects to allow for more frequent monitoring. The approval letter or related documentation will indicate this information.

Because IRB review of research studies must be completed at least annually, investigators should plan ahead to meet required continuing review dates. **Please submit complete continuing review documentation at least 45 days prior to the expiration date to guard against a lapse in IRB approval.** The signed continuing review facepages and any other required hard copies must be received by the HRPP office before the continuing review process can begin.

As a courtesy, automated continuing review reminders can be set-up by PIs at various intervals (75 days, 45 days, 30 days, for example) on the website at <https://irb.ucsd.edu>. However, as these are automated electronic messages based on data entered, and the HRPP cannot anticipate which type of software programs (such as spam-blockers or anti-virus software) may block receipt of the messages, **PI's are required to not rely upon notification, but have internal mechanisms which track continuing review submission times.** Ultimately, it is the PI's responsibility to initiate a continuing review application, allowing sufficient time for the review and re-approval process to be completed before the current approval expires.

Continuing review is required even if no changes are made, or if the only study activity is participant follow-up, and even if the only study activity is data analysis.

What happens if there is a lapse in IRB approval?

If the IRB has not reviewed and approved a research study by the study expiration date, **all research activities must stop**. This includes the following:

All research-related interventions or interactions with currently enrolled subjects (unless the IRB finds that it is in the best interests of the individual subjects to continue participating in the research interventions or interactions;*) recruitment and informed consent procedures; and continued collection and/or analysis of data/information.

**Exception:* Research-related interventions or interactions with enrolled subjects may continue if the IRB determines that stopping the research would jeopardize the rights or welfare of current subjects. The IRB will decide which subjects should continue receiving the intervention during the lapse in approval. A request for such an exception must be submitted in writing to the attention of the IRB Chair by the Principal Investigator. If any project activity—even activity required for participant safety—occurs or continues after the expiration date, the investigator is out of compliance with both federal regulations and university policy. Retrospective approval for work done after the expiration date cannot be granted. For studies involving the VA, additional conditions apply. See VA Handbook 1200.5, Investigator Responsibilities, http://www1.va.gov/vhapublications/ViewPublication.asp?pub_ID=418.

Amendment/revision to an IRB approved study

IRB approval is required before implementing any changes in the approved research plan, consent documents, recruitment materials, or other study-related documents. Please see Amendment Fact Sheet at <http://irb.ucsd.edu/amendmodchg.pdf> for submission guidance.

Adverse Event and Unanticipated Problems Reporting

All problems having to do with subject safety must be reported to the IRB within ten working days. All deaths, whether or not they are directly related to study procedures, must be reported. For adverse events, please utilize the form found at http://irb.ucsd.edu/AE_biomedical_plain.doc. For deviations and other reports, a cover letter and any supplemental information appropriate to the review should be provided. Please see IRB Guidelines for more information at <https://irb.ucsd.edu>.

Changes in financial Interest or Conflict of Interest (COI) disclosure

Any changes in the financial relationship between the study sponsor and any of the investigators on the study and/or any new potential conflicts of interest must be reported immediately to the Independent Review Committee via the Conflict of Interest Office. If these changes affect the conduct of the study or result in a change in the required wording of the approved consent form, then these changes must also be submitted as an amendment request.

Chien, Kathleen

From: George Church [gmc@harvard.edu]
Sent: Wednesday, July 21, 2010 1:11 PM
To: Yveta_masarova@genetics.med.harvard.edu
Cc: Aach, John Dennis (forwarding); 'Jason Bobe'
Subject: CITI completed.

<https://www.citiprogram.org/members/learnersII/crbystage.asp?strKeyID=75C94B20-E463-4CBD-8177-3290F9538792-6342594&gradebook=31782>

CITI Collaborative Institutional Training Initiative

Biomedical Research Investigators Curriculum Completion Report Printed on 7/21/2010

Learner: george church (username: geochurch)
Institution: Harvard Medical School and Harvard School of Dental Medicine
Contact Information Department: Genetics
 Email: gmc@harvard.edu

Biomedical Research Investigators:

Stage 1. Basic Course Passed on 07/21/10 (Ref # 4680017)

Required Modules	Date Completed	Score
Belmont Report and CITI Course Introduction	07/21/10	3/3 (100%)
History and Ethical Principles	07/21/10	6/7 (86%)
Basic Institutional Review Board (IRB) Regulations and Review Process	07/21/10	4/5 (80%)
Informed Consent	07/21/10	4/4 (100%)
Social and Behavioral Research for Biomedical Researchers	07/21/10	4/4 (100%)
Records-Based Research	07/21/10	1/2 (50%)
Research With Protected Populations - Vulnerable Subjects: An Overview	07/21/10	4/4 (100%)
Conflicts of Interest in Research Involving Human Subjects	07/21/10	1/2 (50%)
Harvard Faculty of Medicine	07/21/10	no quiz
Elective Modules	Date Completed	Score
Genetic Research in Human Populations	07/21/10	2/2 (100%)
Vulnerable Subjects - Research Involving Minors	07/21/10	2/3 (67%)
Vulnerable Subjects - Research Involving Pregnant Women and Fetuses in Utero	07/21/10	3/3 (100%)
International Research	07/21/10	1/1 (100%)
Group Harms: Research With Culturally or Medically Vulnerable Groups	07/21/10	2/3 (67%)

For this Completion Report to be valid, the learner listed above must be

affiliated with a CITI participating institution. Falsified information and unauthorized use of the CITI course site is unethical, and may be considered scientific misconduct by your institution.

Paul Braunschweiger Ph.D.
Professor, University of Miami
Director Office of Research Education
CITI Course Coordinator

Return

-- George

<http://arep.med.harvard.edu/gmc>

NRB room 238, <http://maps.google.com/maps?q=77+Ave+Louis+Pasteur>

All appointments must be confirmed with: ymasarova@genetics.med.harvard.edu

CITI Collaborative Institutional Training Initiative

Biomedical Research - Basic/Refresher Curriculum Completion Report Printed on 9/14/2010

Learner: Kun Zhang (username: kun.zhang.ucsd)

Institution: University of California, San Diego

Contact Information 9500 Gilman Drive

MC0412

La Jolla, CA 92093-0412 USA

Department: Bioengineering

Phone: 858-822-7876

Email: kzhang@bioeng.ucsd.edu

Biomedical Research - Basic/Refresher:

Stage 1. Basic Course Passed on 09/14/10 (Ref # 4936774)

Required Modules	Date Completed	Score
Belmont Report and CITI Course Introduction	09/13/10	3/3 (100%)
History and Ethical Principles	09/13/10	7/7 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process	09/13/10	5/5 (100%)
Informed Consent	09/13/10	4/4 (100%)
Social and Behavioral Research for Biomedical Researchers	09/13/10	2/4 (50%)
Records-Based Research	09/13/10	2/2 (100%)
Genetic Research in Human Populations	09/13/10	2/2 (100%)
Research With Protected Populations - Vulnerable Subjects: An Overview	09/14/10	4/4 (100%)
Vulnerable Subjects - Research with Prisoners	09/14/10	4/4 (100%)
Vulnerable Subjects - Research Involving Minors	09/14/10	3/3 (100%)
Vulnerable Subjects - Research Involving Pregnant Women and Fetuses in Utero	09/14/10	1/3 (33%)
Group Harms: Research With Culturally or Medically Vulnerable Groups	09/14/10	3/3 (100%)
FDA-Regulated Research	09/14/10	5/5 (100%)
HIPAA and Human Subjects Research	09/14/10	2/2 (100%)
Conflicts of Interest in Research Involving Human Subjects	09/14/10	2/2 (100%)
University of California, San Diego	09/14/10	no quiz

For this Completion Report to be valid, the learner listed above must be affiliated with a CITI participating institution. Falsified information and unauthorized use of the CITI course site is unethical, and may be considered scientific misconduct by your institution.

Paul Braunschweiger Ph.D.
Professor, University of Miami
Director Office of Research Education
CITI Course Coordinator

[Return](#)

CITI Collaborative Institutional Training Initiative

Human Research Curriculum Completion Report Printed on

Learner: George Daley (username: GQDaley)

Institution: Children's Hospital of Boston

Contact Information Children's Hospital

300 Longwood Ave

Boston, MA 02115 USA

Department: hematology/oncology

Phone: 617-919-2013

Email: george.daley@childrens.harvard.edu

Group 1.Biomedical: Biomedical Investigators and Key Personnel Who Interact With Subjects

Stage 0. Refresher 3 Course Passed on 12/19/08 (Ref # 2378841)

Required Modules	Date Completed	Score
Refresher Course 101 Introduction	12/19/08	no quiz
101 Refresher Course - History and Ethics	12/19/08	1/2 (50%)
101 Refresher Course - Regulations and Process	12/19/08	3/3 (100%)
101 Refresher Course - Informed Consent	12/19/08	2/2 (100%)
101 Refresher Course - Social and Behavioral Research	12/19/08	2/2 (100%)
101 Refresher Course - Records Based Research	12/19/08	no quiz
101 Refresher Course - Genetics Research	12/19/08	2/2 (100%)
101 Refresher Course - An Overview of Research with Vulnerable Subjects	12/19/08	2/2 (100%)
101 Refresher Course - Research with Vulnerable Populations - Prisoners	12/19/08	2/2 (100%)
101 Refresher Course - Research with Vulnerable Populations - Minors	12/19/08	4/4 (100%)
101 Refresher Course - Pregnant Women and Fetuses	12/19/08	3/3 (100%)
101 Refresher Course - FDA Regulated Research and Conference on Harmonization	12/19/08	3/3 (100%)
101 Refresher Course - Conducting human subjects Reserch at the VA	12/19/08	no quiz
101 Refresher Course - Complete the course	12/19/08	no quiz
Children's Hospital of Boston	12/19/08	no quiz

For this Completion Report to be valid, the learner listed above must be affiliated with a CITI participating institution. Falsified information and

unauthorized use of the CITI course site is unethical, and may be considered scientific misconduct by your institution.

Paul Braunschweiger Ph.D.
Professor, University of Miami
Director Office of Research Education
CITI Course Coordinator

[Return](#)

CITI Collaborative Institutional Training Initiative

Human Research Curriculum Completion Report Printed on 9/30/2010

Learner: Jae Joung (username: jj043)
Institution: Massachusetts General Hospital
Contact Information: 149 13th Street, Room 6019
 Molecular Pathology Unit
 Charlestown, MA 02129 USA
 Department: Pathology
 Phone: 617-726-9462
 Email: jjoung@partners.org

Biomedical Research Investigators and Key Personnel:

Stage 1. Basic Course Passed on 08/08/10 (Ref # 4740713)

Required Modules	Date Completed	Score
Introduction	08/07/10	no quiz
History and Ethical Principles	08/07/10	7/7 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process	08/07/10	5/5 (100%)
Informed Consent	08/07/10	4/4 (100%)
Social and Behavioral Research for Biomedical Researchers	08/07/10	4/4 (100%)
Records-Based Research	08/07/10	1/2 (50%)
Genetic Research in Human Populations	08/08/10	2/2 (100%)
Research With Protected Populations - Vulnerable Subjects: An Overview	08/08/10	4/4 (100%)
Vulnerable Subjects - Research Involving Minors	08/08/10	3/3 (100%)
Vulnerable Subjects - Research Involving Pregnant Women and Fetuses in Utero	08/08/10	3/3 (100%)
Group Harms: Research With Culturally or Medically Vulnerable Groups	08/08/10	3/3 (100%)
FDA-Regulated Research	08/08/10	5/5 (100%)
Research and HIPAA Privacy Protections	08/08/10	2/2 (100%)
Conflicts of Interest in Research Involving Human Subjects	08/08/10	1/2 (50%)

For this Completion Report to be valid, the learner listed above must be affiliated with a CITI participating institution. Falsified information and unauthorized use of the CITI course site is unethical, and may be considered scientific misconduct by your institution.

Paul Braunschweiger Ph.D.
 Professor, University of Miami
 Director Office of Research Education
 CITI Course Coordinator

Return

Form Approved Through 11/30/2010

OMB No. 0925-0001

Department of Health and Human Services Public Health Services Grant Application <i>Do not exceed character length restrictions indicated.</i>		LEAVE BLANK—FOR PHS USE ONLY. Type Activity Number Review Group Formerly Council/Board (Month, Year) Date Received	
1. TITLE OF PROJECT (Do not exceed 81 characters, including spaces and punctuation.) "Causal Transcriptional Consequences of Human Genetic Variation"			
2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION <input type="checkbox"/> NO <input checked="" type="checkbox"/> YES Number: PAR-08-094 Title: Centers of Excellence in Genomics Science			
3. PROGRAM DIRECTOR/PRINCIPAL INVESTIGATOR		New Investigator <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
3a. NAME (Last, first, middle) Church, George M.		3b. DEGREE(S) PhD.	
3c. POSITION TITLE Professor		3h. eRA Commons User Name eRA Commons User Name	
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Genetics		3d. MAILING ADDRESS (Street, city, state, zip code) Harvard Medical School NRB 238 77 Avenue Louis Pasteur Boston, MA 02115	
3f. MAJOR SUBDIVISION School of Medicine		E-MAIL ADDRESS: gmc@harvard.edu	
3g. TELEPHONE AND FAX (Area code, number and extension) TEL: 617-432-7562 FAX: 617-432-6513			
4. HUMAN SUBJECTS RESEARCH <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes		4a. Research Exempt If "Yes," Exemption No. <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
4b. Federal-Wide Assurance No. FWA00007071		4c. Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
		4d. NIH-defined Phase III Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
5. VERTEBRATE ANIMALS <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		5a. Animal Welfare Assurance No.	
6. DATES OF PROPOSED PERIOD OF SUPPORT (month, day, year—MM/DD/YY) From 9/13/10 Through 7/31/15		7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD 7a. Direct Costs (\$) 2,954,202 7b. Total Costs (\$) 4,273,088	
		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT 8a. Direct Costs (\$) 13,084,559 8b. Total Costs (\$) 19,585,426	
9. APPLICANT ORGANIZATION Name President and Fellows of Harvard College Address Harvard Medical School Sponsored Programs Administration 25 Shattuck Street Room 509 Boston, MA 02115		10. TYPE OF ORGANIZATION Public: → <input type="checkbox"/> Federal <input type="checkbox"/> State <input type="checkbox"/> Local Private: → <input checked="" type="checkbox"/> Private Nonprofit For-profit: → <input type="checkbox"/> General <input type="checkbox"/> Small Business <input type="checkbox"/> Woman-owned <input type="checkbox"/> Socially and Economically Disadvantaged	
		11. ENTITY IDENTIFICATION NUMBER 1042103580C5 DUNS NO. 047006379 Cong. District MA-008	
12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE Name Deborah Good Title Assoc Dir - Sponsored Programs Admin Address Harvard Medical School 25 Shattuck Street Boston, MA 02115 Tel: 617-432-1596 FAX: 617-432-2651 E-Mail: spa_award@hms.harvard.edu		13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION Name Deborah Good Title Associate Dir - Sponsored Programs Adm. Address Harvard Medical School 25 Shattuck Street Boston, MA 02115 Tel: 617-432-1596 FAX: 617-432-2651 E-Mail: spa_award@hms.harvard.edu	
14. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.		SIGNATURE OF OFFICIAL NAMED IN 13. (In ink. "Per" signature not acceptable.) Signature	

DATE

10/21/10

UCSD # 2009-3915

Form Approved Through 6/30/2012

OMB No. 0925-0001

Department of Health and Human Services Public Health Services Grant Application <i>Do not exceed character length restrictions indicated.</i>		LEAVE BLANK—FOR PHS USE ONLY. Type _____ Activity _____ Number _____ Review Group _____ Formerly _____ Council/Board (Month, Year) _____ Date Received _____	
1. TITLE OF PROJECT (<i>Do not exceed 81 characters, including spaces and punctuation.</i>) Casual Transcriptional Consequences of Human Genetic Variation			
2. RESPONSE TO SPECIFIC REQUEST FOR APPLICATIONS OR PROGRAM ANNOUNCEMENT OR SOLICITATION <input checked="" type="checkbox"/> NO <input type="checkbox"/> YES (If "Yes," state number and title) Number: PAR-08-094 Title: Centers of Excellence in Genomics Science			
3. PROGRAM DIRECTOR/PRINCIPAL INVESTIGATOR			
3a. NAME (Last, first, middle) Zhang, Kun		3b. DEGREE(S) Ph.D.	
		3h. eRA Commons User Name eRA Commons User Name	
3c. POSITION TITLE Assistant Professor		3d. MAILING ADDRESS (<i>Street, city, state, zip code</i>) University of California, San Diego Bioengineering 9500 Gilman Drive, MC 0412 La Jolla, CA 92093-0412	
3e. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Bioengineering			
3f. MAJOR SUBDIVISION General Campus			
3g. TELEPHONE AND FAX (<i>Area code, number and extension</i>) TEL: 858-822-7876 FAX: 858-534-5722		E-MAIL ADDRESS: kzhang@ucsd.edu	
4. HUMAN SUBJECTS RESEARCH No <input type="checkbox"/> Yes <input checked="" type="checkbox"/> IRB Protocol# 071575		4a. Research Exempt <input checked="" type="checkbox"/> No Yes	
4b. Federal-Wide Assurance No. FWA00004495		4c. Clinical Trial <input checked="" type="checkbox"/> No Yes	
		4d. NIH-defined Phase III Clinical Trial <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes	
5. VERTEBRATE ANIMALS <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes		5a. Animal Welfare Assurance No. A3033-01	
6. DATES OF PROPOSED PERIOD OF SUPPORT (month, day, year—MM/DD/YY) From 09/13/10 Through 07/31/15		7. COSTS REQUESTED FOR INITIAL BUDGET PERIOD 7a. Direct Costs (\$) \$93,373	
		7b. Total Costs (\$) \$144,261	
		8. COSTS REQUESTED FOR PROPOSED PERIOD OF SUPPORT 8a. Direct Costs (\$) \$495,729	
		8b. Total Costs (\$) \$765,901	
9. APPLICANT ORGANIZATION Name The Regents of the University of California Address University of California, San Diego - UCSD 9500 Gilman Drive, 0934 La Jolla, California 92093-0934		10. TYPE OF ORGANIZATION Public: → <input type="checkbox"/> Federal <input checked="" type="checkbox"/> State <input type="checkbox"/> Local Private: → <input type="checkbox"/> Private Nonprofit For-profit: → <input type="checkbox"/> General <input type="checkbox"/> Small Business <input type="checkbox"/> Woman-owned <input type="checkbox"/> Socially and Economically Disadvantaged	
		11. ENTITY IDENTIFICATION NUMBER 1956006144A1 DUNS NO 80-435-5790 Cong. District 53	
12. ADMINISTRATIVE OFFICIAL TO BE NOTIFIED IF AWARD IS MADE Name Pamela Alexander Title Contract and Grant Officer Address UCSD / OCGA 9500 Gilman Drive, 0934 La Jolla, California 92093-0934 Tel: 858-534-0240 FAX: 858-534-0280 E-Mail: pjalexander@ucsd.edu		13. OFFICIAL SIGNING FOR APPLICANT ORGANIZATION Name Pamela Alexander Title Contract and Grant Officer Address UCSD / OCGA 9500 Gilman Drive, 0934 La Jolla, California 92093-0934 Tel: 858-534-0240 FAX: 858-534-0280 E-Mail: pjalexander@ucsd.edu	
14. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.		SIGNATURE OF OFFICIAL NAMED IN 13. (In ink. "Per" signature not acceptable.) Signature _____ DATE 10/15/10	

Consultants: \$83,000/yr

Consultants will be engaged for one year terms and evaluated for performance. Expectations of renewal of consultant contracts as described individually below, subject to performance.

Robert C. Green, MD, MPH, \$15,000/yr

Robert C. Green, MD, MPH (consultant)

Robert C. Green, MD, MPH is currently Professor of Neurology, Epidemiology and Genetics at Boston University School of Medicine. He is a physician, clinical trialist and epidemiologist with over twenty years of research experience in health sciences and has been one of the first researchers to direct clinical trials in translational genomics. He serves as Principal Investigator and Director of the REVEAL Study (Risk Evaluation and Education for Alzheimer's Disease), a multi-center project funded by the National Human Genome Research Institute and the National Institute on Aging to understand the impact of genetic risk assessment and APOE disclosure for individuals at risk for Alzheimer's disease. As a consultant on this project, he will assist in the training and mentoring of research fellows and advise on any issues related to human studies that may arise. He will also be involved in the design and implementation of the strategy to recruit, educate and monitor the involvement of minority students as part of the Minority Action Plan (MAP) Portal. Dr. Green will spend approximately EFFORT on this project at a rate of Consultant Info (\$15,000 per year).

Joseph V. Thakuria, MD, \$10,000/yr

Dr. Thakuria is a clinical geneticist at Massachusetts General Hospital who has worked with the Personal Genome Project (PGP) in several capacities: as general medical advisor and strategist; to screen, interview, and advise on selection of candidate applicants; and clinically interpret findings as part of PGP research. Dr. Thakuria will assist the CEGS in identifying and prioritizing genes and variations that may be candidates for CEGS Aim 1 research, and in interpreting the medical and clinical implications of CEGS research and findings. As these will be ongoing areas of research, the expectation is that this role will be continued, subject to performance.

Joe Thakuria will provide consulting work of EFFORT @ Consultant Info total of \$840/month (10,000/year)

Jeantine Lunshof, PhD, \$7,000/yr

Jeantine E. Lunshof, PhD, philosopher and bioethicist, is Assistant Professor, Institute for Public Health Genomics, Faculty Health, Medicine & Life Sciences, Maastricht University, NL and is also a Research Associate, Dept. of Molecular Cell Physiology, VU University Amsterdam, NL, and the Netherlands Institute for Systems Biology (NISB). Dr. Lunshof is developing a research program on the implications of systems biology for normative theory, disease classification, and public health policy, and will assist Dr. Thakuria in prioritizing genes and variations for CEGS Aim 1 research and will interpret the implications of CEGS research and findings in these broader areas. As with Dr. Thakuria, the expectation is that this role will be continued, subject to performance.

Jeantine will provide consulting work of EFFORT @ a rate of Consultant Info total of (\$7,000/year)

Jason Bobe, \$27,000/yr

Jason Bobe, Director of Community for the Personal Genome Project, will assist in the design, development, and distribution of informational and educational materials produced by the CEGS center in support of its Training and Data and Materials Dissemination components. Although these are ongoing functions of the Center, design and development of these informational materials may not be required for its full term and the ongoing need for this role will be examined year by year.

Jason Bobe will work EFFORT @ a rate of Consultant Info total of (\$27,000/year)

JHV Consulting (Ward Vandewege) and Tom Clegg, \$12,000/yr each

Ward Vanderwege has worked with Church Lab post-doc Alexander Wait Zaranek on design and setup of hardware, software, and network configurations for a 96-node data-intensive, high performance computing cluster used by the Church Lab. Tom Clegg has also worked with Alexander Wait on development and maintenance of software infrastructure for storage and processing of CEBS DNA sequence, transcriptome data, and other data; development and maintenance of analysis tools for researchers; system administration for research systems and publicly accessible web services; and technical support for cloud computing applications. Ward Vanderwege and Tom Clegg will similarly work on these complementary aspects of development and implementation of the computational infrastructure required for management and bioinformatic analysis of the genome scale sequence, expression, association, annotation, and other large scale data required for Aim 1, as well as the cell image and single cell transcriptome data developed in Aims 3 and 4. Aim 1 data processing requirements are expected to be heaviest in the beginning and then level off, while Aim 3 single cell transcriptome and Aim 4 image data will likely be light to start and increase later. The expectation is therefore that different levels of the complementary services provided by Ward and Tom will be required in different years, and their roles and will therefore be examined and adjusted accordingly year by year.

JHV consulting and Tom Clegg will provide consulting work of

EFFORT

@

Consultant Info

Consultant Info in total of \$1,000/month (12,000/year)

From: [Chien, Kathleen](#)
To: [Schloss, Jeff \(NIH/NHGRI\) \[E\]](#)
Cc: [Aach, John Dennis \(forwarding\); gmc@harvard.edu; "Yveta Masarova"; Chen, Debbie P \(NIH/NHGRI\) \[E\]](#)
Subject: Church_P50HG005550_JIT
Date: Thursday, June 17, 2010 8:37:34 AM
Attachments: [Church_P50HG005550.pdf](#)

Dear Dr. Schloss,

Attached is the requested Other Support for key personnel under Dr. Church's P50 project. Also attached is the update equipment information. Please let us know if you need any additional information.

Thank you.

Kathleen Chien
Senior Sponsored Research Administrator
Sponsored Programs Administration
Harvard Medical School
Gordon Hall, Suite 509
25 Shattuck St.
Boston, MA 02115
617-432-5801
fax: 617-432-2651

Specialized Equipment: \$499,000/ yr 1**CTCHGV Special Equipment**

ITEM	estimate
MAGE device for regulatory region alterations (Aim 1)	\$80,000
Stem cells: Inverted phase contrast & fluorescence microscope with video output for cell imaging (Aim 2)	\$15,000
Stem cells: Single chamber Nucleofector (Aim 2)	\$12,000
Stem cells: Liquid nitrogen storage tank (Aim 2)	\$8,000
Total Internal Reflection Fluorescence (TIRF) microscope objective (Aim 3)	\$15,000
Hamamatsu CCD for single molecule imaging (Aim 3)	\$37,000
Biolist gene gun (Aim 3)	\$27,000
Akta FPLC (GE) (Aim 4.2)	\$40,000
Polonator for synthetic biology and cell sorting (Aims 4.1, 4.3)	\$167,000
Agilent Bio Analyzer (all Aims)	\$25,000
Computers for four CTCHGV post-docs (all Aims)	\$10,000
Shaker (all Aims)	\$15,000
8 high performance compute nodes for CTCHGV to be added to Harvard cluster (based on specs: IBM System x3550, Dual core/Dual CPU Intel Xeon EM64T CPUs (4 cores) , 8 GB RAM, 2.73 GB SAS disks (computational support for all Aims)	\$48,000
TOTAL	\$499,000

CTCHGV Special Equipment budget

PHS 398/2590 OTHER SUPPORT

Provide active support for all key personnel. Other Support includes all financial resources, whether Federal, non-Federal, commercial or institutional, available in direct support of an individual's research endeavors, including but not limited to research grants, cooperative agreements, contracts, and/or institutional awards. Training awards, prizes, or gifts do not need to be included.

GEORGE M. CHURCH, PhD.**ACTIVE:****DE-FG02-02ER63445 (GTL)**

2/1/2003 – 11/30/2011

EFFORT

DOE-GTL

\$1,235,477 direct cost/yr

Principal Investigator: George Church

Title: Microbial Ecology, Proteogenomics & Computational Optima.

Project studies proteomics and cell models for *Prochlorococcus* and *Pseudomonas***SA5283-11210 (NSF)**

7/1/2006 – 6/30/2011

EFFORT

NSF-(SynBERC)

\$146,137 direct cost/yr

Principal Investigator: Jay Keasling (UC Berkeley)

Title: Synthetic Biology Engineering Research Center

Our role is to develop synthetic bacterial genome "chasses" for safe use in mammals

W911NF-08-1-0254 (DARPA)

DARPA

6/27/08 – 10/2010

EFFORT

Phase IIB

2/1/10-10/31/10

Principal Investigator: Neil Gershenfeld (MIT)

\$63,582 direct cost/yr

Title: Milli-Biology: Programmed Assembly of Engineered Materials

RO1 HL 094963-01 (NHLBI)

NIH - NHLBI

9/30/2008-6/30/2011 (NCE)

EFFORT

Principal Investigator: George Church

Church: \$369,700 direct cost/yr

Subcontract: J. Seidman (HMS)

Seidman: \$442,225 direct cost/yr

Subcontract: K Zhang (UCSD)

Zhang: \$438,451 direct cost/yr

Title: Targeted 2nd generation sequencing in phenotyped Framingham & PGP populations

Project Goal: We propose to develop, demonstrate, and validate a pipeline for high-throughput, low-cost targeted resequencing of all human exons based on next-generation (gen2) sequencing techniques in support of the long term goal of enabling sequencing to be used routinely to characterize genotypes and genetic variation in genome-wide medical targets for large populations of individuals

Private Source

11/01/08-10/30/12

EFFORT

Principal Investigator: George Church

\$150,000 direct cost/yr

Private Source

The main goal of this project is the identification and characterization of naked mole-rat genes that contributed to the evolution of a long lifespan in this species.

RC2 HG005592 (NHGRI)

10/01/09-09/30/11

EFFORT

NIH-NHGRI - Halcyon

Church: \$109,958 direct cost/yr2

PI: George Church

Sub: Halcyon: 1,000,000 /yr

Sub: Halcyon

Title: Development of Electron Microscopy-based Nucleic Acid Polymer Sequencing

Project: We aim to provide a comprehensive foundation for development of an ultra-low-cost, ultra-fast nucleic acid polymer sequencing technology based on single-atom resolution transmission electron microscopy (TEM) of heavy atom-labeled nucleic acid polymers.

RC2 HL102815 (NHLBI)

9/30/09-9/29/11

EFFORT

NIH- NHLBI

Sub: Church: \$76,395 direct cost/yr

PI: George Daley (Children's Hosp)

Sub: George Church

Title: Comparative phenotypic, functional, and molecular analysis of ESC and iPSC

Role: to develop high-throughput and genome-wide profiling methods for characterizing human induced pluripotent stem cells, relevant but not limited to their epigenetic changes, alterations in transcript isoforms and non-coding RNA molecules along with single cell reprogramming and analysis.

P50 HG003170 (CEGS supplement)

7/1/09-6/30/10

EFFORT

NIH- NHGRI

Church: \$581,282 direct cost/yr

PI: George Church

Title: Molecular and Genomic Imaging Center

Role: The goal of this request for supplemental funding for our Molecular and Genomic Imaging CEGS (MGIC) is further development of MGIC initiatives focused on single cell and splice isoform analysis along with enabling MGIC technology.

ONRBAA09-001

Office of Naval Research

4/1/10-3/31/13

EFFORT

PI: George Church

\$45,551 direct cost/yr

Title: Multiplexed Pathway and Organism Engineering.

RC1 HG005482

NIH/NCRR

9/22/09-6/30/11

EFFORT

PI: Peter Park

\$25,756 direct cost/yr

Sub: George Church

Title: Statistical Methods for Estimation of Copy Number from Next – Generation Sequencing

Overlap:

None

Pending:

Pending Support

JOUNG, J.K.ACTIVE:

R01 GM069906-05 (Joung)(no-cost extension) 08/01/04-07/31/10 EFFORT
 NIH/NIGMS \$177,066
 Studies of NRSF/REST zinc finger-DNA interactions
 Massachusetts General Hospital
 This grant is aimed at studying protein-DNA interactions by the NRSF/REST transcription factor.

R24 GM078369-03 (Joung) 03/01/07-02/28/11 EFFORT
 NIH/NIGMS \$83,426
 DNA-binding specificities of Cys2His2 zinc fingers
 Massachusetts General Hospital
 This grant award is aimed at developing a probabilistic recognition code for Cys2His2 zinc fingers.

R01 GM088040-01 (Joung/Peterson) 08/01/09-07/31/13 EFFORT
 NIH/NIGMS \$202,791
 Advancing Zinc Finger Nucleases for Targeted Genome Manipulation
 Massachusetts General Hospital
 This grant is aimed at developing zinc finger nuclease technology for use in zebrafish.

ZINC-HUBS 201249 (Isalan) 03/01/09-02/28/11 EFFORT
 European Research Council \$55,000
 Engineering zinc fingers to target cancer hub genes
 Centre for Genomic Regulation (MGH subcontract)
 This subcontract supports engineering customized zinc finger domains for use in system biology experiments

DBI-0923827 (Voytas) 09/15/09 -08/31/13 EFFORT
 NSF \$109,972
 Precision Engineering of Plant Genomes Using Zinc Finger Nucleases
 University of Minnesota (MGH subcontract)
 This subcontract supports the development of zinc finger nuclease reagents for modifying genes in rice.

R01 CA057683-16S1 (Louis) 07/01/09-06/30/10 0.60 calendar
 NIH/NCI \$35,000
 Toward a molecular classification of human gliomas
 Massachusetts General Hospital
 This supplement supports the development of a combinatorial zinc-finger transcription factor library for studying drug resistance in gliomas.

RC2 HL101553-01 (Orkin) 09/30/09-09/29/11 EFFORT
 NIH/NHLBI \$88,826
 Extending Gwas in the BCL11 locus to novel therapeutics for HbF induction
 Children's Hospital – Boston (MGH subcontract)
 This subcontract is aimed at supporting the development of zinc finger nucleases targeted to *BCL11*.

PENDING:

Pending Support

VIZ00099725

DALEY, GEORGE, Q.

ACTIVE

U01HL100001 (Daley)

9/30/09-6/30/16

EFFORT

\$250,000 DC Year 1

NHLBI Progenitor Cell Biology Consortium Research Hubs (U01)

Human Pluripotent Stem Cell and Progenitor Models of Cardiac and Blood Diseases

This grant aims to generate human iPS cells from patients with specific genetic and acquired blood disorders, to explore the blood phenotypes of these iPS cells, to investigate methods for gene repair, and to pursue chemical and genetic screening to identify novel small molecules and genetic pathways to ameliorate the disease phenotypes *in vitro*. Co-PIs are Len Zon and Stu Orkin. Cardiac hub PI: Ken Chien, MGH.

NIH 1RC2HL102815 (Daley)

9/30/2009-9/29/2011

EFFORT

NIH/NIDDK

\$304,193 DC Year 1

Comparative phenotypic, functional, and molecular analysis of ESC and iPSC

The chief goal of this multi-PI, multi-institutional collaborative project is to catalogue the epigenetic similarities and differences among various types of pluripotent stem cells.

NIH U01 DK072473 (Melton)

9/1/2009-8/31/2011

EFFORT

\$189,250 DC Year 1

Differentiation of iPS cells from type 1 diabetics into hematopoietic stem cells

The aims of this sub-contract include exploration of hematopoietic and pancreatic differentiation from iPS cells generated from patients with type 1 diabetes, with the ultimate goal of combined hematopoietic/ pancreatic models of human cell transplants in immune-deficient mice.

Private Source

10/1/07-9/30/10

EFFORT

\$200,000 DC Year 1

Private Source

This grant evaluates new kinase inhibitors for activity against resistant cases of CML, and investigates new mechanisms of disease progression.

Private Source

6/1/09-5/31/11

EFFORT

~\$93,500 (100,000 CHF) DC Year 1

Private Source

This grant aims to explore hematopoietic gene expression in genetically abnormal mouse and human pluripotent cells, and to interrogate the functional significance of aplastic anemia candidate genes.

Private Source

7/1/09-6/30/11

EFFORT

\$100,000 DC Year 1

Private Source

This grant will test the hypothesis that modulation of Lin28 and Lin28 targets will influence formation of germ cell tumors (GCTs), and will define Lin28 role in primary human GCT biopsies.

NIH RO1 DK70055 (Daley)

7/1/2009-6/30/11

EFFORT

NIH/NIDDK

\$28,000 DC Year 1

Administrative Supplement – Summer Students

Murine Models for Regenerative Medicine

This administrative supplement will fund 6 summer students and a portion of a Scientific Coordinator to work on projects related to the parent RO1. This administrative supplement overlaps with parent RO1 DK70055.

Private Source

1/15/08-7/31/13

By policies of the Private Source funding supports pilot and innovative projects for which traditional funding would be exceedingly difficult to obtain, projects that are not covered by existing NIH grants, and research thrusts that extend beyond the allotments in NIH-funded projects. Private Source supports only part of the total laboratory's budget (in our case, approximately 25%). Accordingly, non-Private Source support is critical for the success of our research.

PENDING

Pending Support

Program Director/Principal Investigator: ZHANG, Kun

For New and Renewal Applications (PHS 398) – DO NOT SUBMIT UNLESS REQUESTED
For Non-competing Progress Reports (PHS 2590) – Submit only Active Support for Key Personnel

PHS 398/2590 OTHER SUPPORT

Provide active support for all key personnel. Other Support includes all financial resources, whether Federal, non-Federal, commercial or institutional, available in direct support of an individual's research endeavors, including but not limited to research grants, cooperative agreements, contracts, and/or institutional awards. Training awards, prizes, or gifts do not need to be included.

Kun Zhang, PhD.

ACTIVE

1R01HG004876-02 (Zhang)
NIH/NHGRI

08/01/09-07/31/10
\$373,866

EFFORT

Title: An integrated lab-on-chip system for genome sequencing of single microbial cells.
The major goals of this project are to develop an integrated genomic platform for genetic analysis and genome sequencing of single microbial cells associated with human.

1R01DA025779-02 (Zhang)
NIH/NIDA

08/01/09-07/31/10
\$202,101

EFFORT

Title: Genome-scale analysis of DNA methylation in CpG islands through bisulfite sequencing.
The major goals of this project are to develop methods for specifically capture an arbitrary subset of genome for DNA methylation analysis, and to apply the methods to characterizing epigenetic changes in the differentiation of stem cells.

1 R01 HL094963-01 (Church)
NIH/NHLBI

07/01/09-06/30/10
\$439,706

EFFORT

Title: Targeted 2nd Generation Sequencing in Phenotyped Framingham & PGP Populations.
The major goals of this project are to develop methods for specifically capture and re-sequence all protein coding regions in the human genome at the cost of approximately \$1,000.

PENDING:

Pending Support

From: Chien, Kathleen
To: Chen, Debbie P (NIH/NHGRI) [E]
Cc: Baker-Webber, Pamela A (forwarding); "Yveta Masarova"
Subject: Church_P50HG0055550
Date: Wednesday, August 11, 2010 10:47:43 AM
Attachments: othersupport-KunZhang-2010.doc

Hi,

Attached is the requested updated Other Support for Dr. Zhang. Please let us know if you need any additional information.

Best,

Kathleen Chien
SSRA, Sponsored Programs Administration
Harvard Medical School
25 Shattuck St., Suite 509
Boston, MA 02115
617-432-5801
kathleen_chien@hms.harvard.edu

Program Director/Principal Investigator: ZHANG, Kun

For New and Renewal Applications (PHS 398) – DO NOT SUBMIT UNLESS REQUESTED
For Non-competing Progress Reports (PHS 2590) – Submit only Active Support for Key Personnel

PHS 398/2590 OTHER SUPPORT

Provide active support for all key personnel. **Other Support includes all financial resources, whether Federal, non-Federal, commercial or institutional, available in direct support of an individual's research endeavors, including but not limited to research grants, cooperative agreements, contracts, and/or institutional awards.** Training awards, prizes, or gifts do not need to be included.

Kun Zhang, PhD.

ACTIVE

1R01HG004876 (Zhang)
NIH/NHGRI

08/01/10-07/31/11
\$373,866

EFFORT

Title: An integrated lab-on-chip system for genome sequencing of single microbial cells.

The major goals of this project are to develop an integrated genomic platform for genetic analysis and genome sequencing of single microbial cells associated with human.

1R01DA025779 (Zhang)
NIH/NIDA

08/01/10-07/31/11
\$202,101

EFFORT

Title: Genome-scale analysis of DNA methylation in CpG islands through bisulfite sequencing.

The major goals of this project are to develop methods for specifically capture an arbitrary subset of genome for DNA methylation analysis, and to apply the methods to characterizing epigenetic changes in the differentiation of stem cells.

1R01 HL094963 (Church)
NIH/NHLBI

07/01/10-06/30/11
No cost extension

EFFORT

Title: Targeted 2nd Generation Sequencing in Phenotyped Framingham & PGP Populations.

The major goals of this project are to develop methods for specifically capture and re-sequence all protein coding regions in the human genome at the cost of approximately \$1,000.

PENDING:

Pending Support

From: [Graham, Bettie \(NIH/NHGRI\) \[E\]](#)
To: [Chen, Debbie P \(NIH/NHGRI\) \[E\]](#)
Cc: [Schloss, Jeff \(NIH/NHGRI\) \[E\]](#)
Subject: FW: 1P50HG005550-01 (CHURCH): revised MAP budget
Date: Tuesday, April 27, 2010 7:02:20 AM
Attachments: [MAP budget revised 4-22-10.pdf](#)

Good Morning,

I asked George to respond to some of the issues in the assessment and a revised budget.

I hope that this budget can be incorporated in the parent grant and that a supplement memo will not be necessary. Please advise because we will need this information for the Mock Pay List meeting on Friday.

Thanks,

Bettie

From: Good, Deborah [mailto:deborah_good@hms.harvard.edu]
Sent: Friday, April 23, 2010 2:24 PM
To: Graham, Bettie (NIH/NHGRI) [E]
Cc: Evans, Kelly A.
Subject: FW: 1P50HG005550-01 (CHURCH): revised MAP budget

Dear Dr. Graham,

Attached is the letter of response to the summary statement from the 3/23/10 review of the subject grant's revised MAP proposal, along with the revised MAP budget and budget justification. Please let us know if you require any additional information.

Sincerely,

Deborah Good
Associate Director
Sponsored Programs Administration
Harvard Medical School
Boston, MA 02115

(617) 432-2911

Harvard Medical School
Department of Genetics
77 Avenue Louis Pasteur
Boston, MA, USA 02115
<http://arep.med.harvard.edu>



April 22, 2010

Bettie Graham, Ph.D.
Associate Director, Division of Extramural Research
National Human Genome Research Institute
National Institutes of Health
5635 Fishers Lane
Room 4076, MSC 9305
Rockville, MD 20892-9305

Dear Bettie,

We are very pleased with the overall positive response of the Special Emphasis Panel's March 23 review of our February 12 revised MAP proposal for our CTCHGV CEGS. However, we recognize the outstanding issues that were noted in their review comments and which we discussed during our phone conversation of April 17. We respond to these concerns here and in the attached revised MAP budget and budget justification, in which those items with budgetary implications are considered in more detail.

Evaluative Info

Page 175 of 906

Withheld pursuant to exemption

Evaluative Info

under the Freedom of Information and Privacy Act

VIZ00099733

Evaluative Info

Sincerely,

Signature

George M. Church,
Professor of Genetics, Harvard Medical School
MIT Health Sciences & Technology
Senior Associate of the Broad Inst. of Harvard & MIT
Director, DOE Biofuels Center, Personal Genome Project
& NIH Center for Excellence in Genomic Science

Budget - MAP y11

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY						FROM 4/1/10	THROUGH 3/31/11	
PERSONNEL (Applicant organization only)		Months Devoted to Project			INST.BASE SALARY	DOLLAR AMOUNT REQUESTED (omit cents)		
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths		SALARY REQUESTED	FRINGE BENEFITS	TOTAL
Church, George M.	PD/PI					0		0
TBH-PostDoc	PostDoc	EFFORT			Institutional Base Salary			
TBH -PostDoc	PostDoc					39,000	10,725	49,725
TBH	MAP Manager					78,462	32,954	111,415
TBN- summer	student					4,000		4,000
TBN- summer	student					4,000		4,000
TBN- summer	student					4,000		4,000
TBN- summer	student					4,000		4,000
TBN- summer	student					4,000		4,000
SUBTOTALS						137,462	43,679	181,140
CONSULTANT COSTS								
EQUIPMENT (Itemize)								
SUPPLIES (Itemize by category)								
Lab supplies: \$ 20,000								
Computers and software : \$ 2,000								
22,000								
TRAVEL								
Travel – to conferences - recruiting PostDocs: \$ 7,500								
Travel- undergrad interns (500x5) : \$2,500								
Travel- Postdoc interviews (1000x2): \$2,000								
12,000								
PATIENT CARE COSTS								
INPATIENT								
OUTPATIENT								
ALTERATIONS AND RENOVATIONS (Itemize by category)								
OTHER EXPENSES (Itemize by category)								
Recruiting costs: advertisement (websites) : 500								
Lunch meetings for MAP students: (4 meetings): 400								
Mentoring event: 600								
GRE courses for undergrad interns: (5x400): 2,000								
Living expenses/housing for undergrad interns (2870x5): 14,350								
17,850								
CONSORTIUM/CONTRACTUAL COSTS						DIRECT COSTS		
SUBTOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD (Item 7a, Face Page)						\$ 232,990		
CONSORTIUM/CONTRACTUAL COSTS						FACILITIES AND ADMINISTRATIVE COSTS		
TOTAL DIRECT COSTS FOR INITIAL BUDGET PERIOD						\$ 232,990		

MAP - Syt Budget

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

**BUDGET FOR ENTIRE PROPOSED PROJECT PERIOD
DIRECT COSTS ONLY**

BUDGET CATEGORY TOTALS		INITIAL BUDGET PERIOD (from Form Page 4)	ADDITIONAL YEARS OF SUPPORT REQUESTED			
			2nd	3rd	4th	5th
PERSONNEL: <i>Salary and fringe benefits. Applicant organization only.</i>		181,140	202,955	208,737	214,399	220,231
CONSULTANT COSTS						
EQUIPMENT						
SUPPLIES		22,000	43,260	44,558	45,895	47,271
TRAVEL		12,000	12,300	12,609	12,927	11,255
PATIENT CARE COSTS	INPATIENT					
	OUTPATIENT					
ALTERATIONS AND RENOVATIONS						
OTHER EXPENSES		17,850	20,622	21,210	21,816	21,878
CONSORTIUM/ CONTRACTUAL COSTS	DIRECT					
SUBTOTAL DIRECT COSTS (Sum = Item 8a, Face Page)		232,990	279,137	287,114	295,037	300,636
CONSORTIUM/ CONTRACTUAL COSTS	F&A					
TOTAL DIRECT COSTS		232,990	279,137	287,114	295,037	300,636
TOTAL DIRECT COSTS FOR ENTIRE PROPOSED PROJECT PERIOD						\$ 1,394,914

JUSTIFICATION. Follow the budget justification instructions exactly. Use continuation pages as needed.

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): George Church

Budget Justification – Harvard

*Revised April 22, 2010 in response to March 23, 2010 review of our Feb 12, 2010 proposal revision.
Revisions are indicated in italics.*

Church lab personnel:**Dr. George Church**

George Church is principal investigator, will direct the research activities of the CEGS and the overall direction and activities of its MAP program.

MAP Manager, TBH

The roles and duties of the MAP Manager are fully documented in the MAP proposal in section C.3.1.i. Briefly, the MAP Manager will share responsibility with Dr. Church and with his support design, direct, and implement a structured set of MAP processes for recruiting underrepresented minority (URM) post-docs and summer undergraduates, establishing them in research projects, providing venues and materials for their training and mentoring, interfacing with other programs and resources for education and career development, and tracking their progress and soliciting feedback from them both during and after their terms with the Center. The MAP Manager will not only oversee the execution of these activities but personally participate in them, including their coordination and administration, and be available personally as mentor. Finally, the MAP Manager will represent the Center CEGS at national NHGRI meetings and coordinate responses to NHGRI initiatives, and at conferences for purposes of recruiting post-docs and undergraduates.

The MAP Manager will be a effort for Year 1 and a effort for Years 2-5. Justification for this arrangement is given in section C.3.1.i.

Evaluative Info

Postdoctoral Fellow, PhD, TBH (two post-docs)

The purpose of the MAP program is to recruit and then train and mentor URM students in the context of providing them research experiences in the genomic science of the Center, in order to assist them in advancing towards careers in biological and biomedical research. We have set a Center MAP goal of recruiting one new URM post-doc a year for two year terms at the Center (see proposal section C.2.2), so that in each year of the Center, there will be two post-docs supported by the MAP, one in their first year, and one in their second. The post-docs will, with support and direction from the PI, develop and perform a research project for the Center in support of the Center's genomic sciences Aims. Post-docs will therefore be efforts, but at two different salary levels, i.e., the incoming post-doc will have a lower salary than the second year post-doc.

In brief, these Aims are: (1) Identify variations in *cis* gene regions that causally affect gene transcription levels by engineering precise modifications to the variations and directly observing whether this alters transcription levels. This Aim also includes very significant technology development for genetic engineering techniques, specifically for Zinc Finger Nuclease (ZFN) and Multiplex Automated Genetic Engineering (MAGE) methods, which must be made to operate precisely and efficiently in human cells. Finally, this Aim also includes technology development that will enable these methods to be applied to single human cells, at a throughput which will enable high throughput study of very large populations of single cells. (2) Extend the methods of Aim 1 to operate in human induced Pluripotent Stem cells (iPS) and, with these methods, explore the causality

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): George Church

of cis gene variations on transcription levels in different human tissues. (3) Develop and demonstrate methods for measuring the transcription levels of hundreds to many thousands of transcripts in single human cells, including structured human tissues. This extends the single cell methods to be used in Aim 1. (4) Develop advanced technologies of wide impact to be used in Aims 1-3 that include: integrated sequencing and synthesis of thousands of large DNA constructs, comprehensive identification of a library of ZFNs capable of targeting any location in the human genome, and development of an instrument capable of arraying thousands of individual human cells that will enable the single cell aspects of Aims 1 and 3, and provide a novel cell sorting method with capabilities beyond Fluorescent Activated Cell Sorting (FACS).

Summer Students: (5)

In support of the MAP program goals above, in addition to the post-docs, we also have set a Center MAP goal of supporting five undergraduate students through a EFFORT summer research experience under direction of a Center post-doc, with support in the form of skills sessions, GRE prep courses, mentoring, and other support and career development events (see proposal section C.2.1). Summer students will be provided funding for room and board and a stipend.

Evaluative Info

Supplies:

General lab supplies: \$20000/year has been allocated for laboratory supplies for each MAP post-doc to pursue the Center research project developed and executed by the post-doc. This is in accord with Church Lab standard practice of estimating laboratory supply costs for post-doc-conducted research, generally.

Evaluative Info

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): George Church

Evaluative Info

Travel-recruiting

Evaluative Info

Travel-candidate interview reimbursement

As part of recruiting, we envision inviting two URM post-doc candidates per year to come to the Center to talk with the PI, to give a lab talk, to get to know other Center investigators, post-docs and students, and get a feel for the institutional setting. We expect that one of these two candidates would become the one new post-doc per year that would join the Center.

Evaluative Info

Recruiting-advertising

This cost includes charges for advertisements on web sites or other venues that will better enable us to recruit talented URM post-docs and undergraduates. For instance, SACNAS charges \$150 for 90 day postings on their web site advertising opportunities for post-doc fellowships and undergraduate programs. Several postings of this sort may be required in the course of recruiting.

Lunch meetings (undergraduates)

As described in the section C.2.1.1 of the proposal, we will hold four 1.5 hour skill sessions per summer for our URM summer undergraduates in which we will provide training and discussion on essential skills for conduct of scientific research. To make these fun and informal, we plan on having them as "pizza lunches," and this will also encourage Center grad students and post-docs to attend so that they can join in these discussions. Thus, we estimate that ~15 people might attend each session.

GRE prep courses (undergraduates)

We commit to sending our five undergraduate interns per summer to GRE prep courses, which cost ~\$600 each.

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): George Church

Evaluative Info

Mentoring event

In section C.3.2 of the proposal, we describe our plans to incentivize and support Center faculty, post-docs, and grad students who act as mentors to MAP URM post-docs and undergraduates, which duties, though intrinsically rewarding, can also require considerable time and attention. This budget item represents the costs of these support activities, including a yearly "mentor dinner" at which mentors can be both recognized for their work and have a forum for discussing mentorship issues.

Fringe Benefits:

Budget Justification - HARVARD

Principal Investigator/Program Director(Last, first, middle): George Church

Faculty: 28% (FY11) and 28.4% (FY12).

Postdoctoral Fellow: 27% (FY10), 27.5% (FY11) and 27.8% (FY12)

Professional: 42% (FY10), 43.7% (FY11), 44.2% (FY12)

Technical Assistant: 54.6% (FY10), 56.6% (FY11) and 57.7% (FY12)

Facilities and Administrative Costs:

Facilities and Administrative costs for Harvard Medical School are calculated at 69.5%.

MAP

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

CHECKLIST

TYPE OF APPLICATION (Check all that apply.)

- ☒ NEW application. (This application is being submitted to the PHS for the first time.)
- ☐ RESUBMISSION of application number: _____
(This application replaces a prior unfunded version of a new, renewal, or revision application.)
- ☐ RENEWAL of grant number: _____
(This application is to extend a funded grant beyond its current project period.)
- ☐ REVISION to grant number: _____
(This application is for additional funds to supplement a currently funded grant.)
- ☐ CHANGE of program director/principal investigator.
Name of former program director/principal investigator: _____
- ☐ CHANGE of Grantee Institution. Name of former institution: _____
- ☐ FOREIGN application ☐ Domestic Grant with foreign involvement List Country(ies) Involved: _____

INVENTIONS AND PATENTS (Renewal appl. only) ☐ No ☒ Yes
If "Yes," ☒ Previously reported ☐ Not previously reported

1. PROGRAM INCOME (See instructions.)

All applications must indicate whether program income is anticipated during the period(s) for which grant support is request. If program income is anticipated, use the format below to reflect the amount and source(s).

Budget Period	Anticipated Amount	Source(s)
	0	

2. ASSURANCES/CERTIFICATIONS (See instructions.)

In signing the application Face Page, the authorized organizational representative agrees to comply with the policies, assurances and/or certifications listed in the application instructions when applicable. Descriptions of individual assurances/certifications are provided in Part III and listed in Part I, 4.1 under Item 14. If unable to certify compliance, where applicable, provide an explanation and place it after this page.

3. FACILITIES AND ADMINSTRATIVE COSTS (F&A)/ INDIRECT COSTS. See specific instructions.

- ☒ DHHS Agreement dated: 04/24/2009 ☐ No Facilities And Administrative Costs Requested
- ☐ DHHS Agreement being negotiated with _____ Regional Office.
- ☐ No DHHS Agreement, but rate established with _____ Date _____

CALCULATION* (The entire grant application, including the Checklist, will be reproduced and provided to peer reviewers as confidential information.)

a. Initial budget period:	Amount of base \$	232,990	x Rate applied	69.5	% = F&A costs	\$	161,928
b. 02 year	Amount of base \$	279,137	x Rate applied	69.5	% = F&A costs	\$	194,000
c. 03 year	Amount of base \$	287,114	x Rate applied	69.5	% = F&A costs	\$	199,544
d. 04 year	Amount of base \$	295,037	x Rate applied	69.5	% = F&A costs	\$	205,051
e. 05 year	Amount of base \$	300,636	x Rate applied	69.5	% = F&A costs	\$	208,942
TOTAL F&A Costs							\$ 969,466

*Check appropriate box(es):

- ☐ Salary and wages base ☒ Modified total direct cost base ☐ Other base (Explain)
- ☐ Off-site, other special rate, or more than one rate involved (Explain)

Explanation (Attach separate sheet, if necessary.):

Indirect cost:

4. DISCLOSURE PERMISSION STATEMENT: If this application does not result in an award, is the Government permitted to disclose the title of your proposed project, and the name, address, telephone number and e-mail address of the official signing for the applicant organization, to organizations that may be interested in contacting you for further information (e.g., possible collaborations, investment)? ☐ Yes ☒ No

FEDERAL FINANCIAL REPORT

FINAL

1. Federal Agency and Organizational Element to Which Report is Submitted NATIONAL HUMAN GENOME RESEARCH INSTITUTE			2. Federal Grant or Other Identifying Number Assigned by Federal Agency 1P50HG005550-1				
3.Recipient Organization (Name and complete address, including Zip code) HARVARD UNIVERSITY (MEDICAL SCHOOL) HARVARD UNIVERSITY MEDICAL SCHOOL CAMPUS 25 Shattuck St. Suite 509 BOSTON MA 02115							
4a. DUNS Number 047006379		4b. EIN 1042103580C5	5. Recipient Account Number or Identifying Number		6. Report Type <input type="checkbox"/> Quarterly <input type="checkbox"/> Semi-Annual <input type="checkbox"/> Annual <input checked="" type="checkbox"/> Final	7. Basis of Accounting <input checked="" type="checkbox"/> Cash <input type="checkbox"/> Accrual	
8. Project/Grant Period From: (Month, Day, Year) 09/13/2010			To: (Month, Day, Year) 07/31/2015			9. Reporting Period End Date (Month, Day, Year) 07/31/2011	
10. Transactions						Cumulative	
(Use lines a-c for single or multiple grant reporting)							
Federal Cash (To report multiple grants, also use FFR Attachment):							
a. Cash Receipts						0.00	
b. Cash Disbursements						0.00	
c. Cash on Hand (line a minus b)						0.00	
(Use lines d-o for single grant reporting)							
Federal Expenditures and Unobligated Balance:							
d. Total Federal funds authorized						4,273,088.00	
e. Federal share of expenditures						3,114,294.13	
f. Federal share of unliquidated obligations						0.00	
g. Total Federal share (sum of lines e and f)						3,114,294.13	
h. Unobligated balance of Federal funds (line d minus g)						1,158,793.87	
Recipient Share:							
i. Total recipient share required						0.00	
j. Recipient share of expenditures						0.00	
k. Remaining recipient share to be provided (line i minus j)						0.00	
Program Income:							
l. Total Federal program income earned						0.00	
m. Program income expended in accordance with the deduction alternative						0.00	
n. Program income expended in accordance with the addition alternative						0.00	
o. Unexpended program income (line l minus line m or line n)						0.00	
11. Indirect Expense	a. Type	b. Rate	c. Period From	Period To	d. Base	e. Amount Charged	f. Federal Share
	Predetermined	69.50	09/13/2010	06/30/2011	1,223,263.06	850,167.83	850,167.82
	Predetermined	69.00	07/01/2011	07/31/2011	223,203.22	154,010.22	154,010.22
				g. Totals:	1,446,466.28	1,004,178.05	1,004,178.04
12. Remarks: Attach any explanations deemed necessary or information required by Federal sponsoring agency in compliance with governing legislation: Accepted By: Richard Berardi Date Report Accepted: 12/28/2011. PI will request a carry forward of the unspent balance under separate cover.							
13. Certification: By signing this report, I certify that it is true, complete, and accurate to the best of my knowledge. I am aware that any false, fictitious, or fraudulent information may subject me to criminal, civil, or administrative penalties. (U.S. Code, Title 218, Section 1001)							
a. Typed or Printed Name and Title of Authorized Certifying Official Lisa M. Ortega Financial Analyst II						c. Telephone (Area code, number and extension) 617-432-6285	
						d. Email address lisa_ortega@harvard.edu	
b. Signature of Authorized Certifying Official						e. Date Report Submitted (Month, Day, Year) 11/29/2011	
						14. Agency use only:	

FEDERAL FINANCIAL REPORT

--	--

Standard Form 425
OMB Approval Number: 0348-0061
Expiration Date: 10/31/2011



SPECIALIZED RESEARCH CENTER
Department of Health and Human Services
National Institutes of Health
NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Issue Date: 11/03/2010



Grant Number: 1P50HG005550-01 REVISED

Principal Investigator(s):
GEORGE M CHURCH, PHD

Project Title: Causal Transcriptional Consequences of Human Genetic Variation

ASSOCIATE DIRECTOR, SPA
HARVARD MEDICAL SCHOOL
25 SHATTUCK STREET
BOSTON, MA 02115

Award e-mailed to: spa_award@hms.harvard.edu

Budget Period: 09/13/2010 – 07/31/2011

Project Period: 09/13/2010 – 07/31/2015

Dear Business Official:

The National Institutes of Health hereby revises this award (see "Award Calculation" in Section I and "Terms and Conditions" in Section III) to HARVARD UNIVERSITY (MEDICAL SCHOOL) in support of the above referenced project. This award is pursuant to the authority of 42 USC 241 42 CFR 52 and is subject to the requirements of this statute and regulation and of other referenced, incorporated or attached terms and conditions.

Acceptance of this award including the "Terms and Conditions" is acknowledged by the grantee when funds are drawn down or otherwise obtained from the grant payment system.

Each publication, press release or other document that cites results from NIH grant-supported research must include an acknowledgment of NIH grant support and disclaimer such as "The project described was supported by Award Number P50HG005550 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health."

Award recipients are required to comply with the NIH Public Access Policy. This includes submission to PubMed Central (PMC), upon acceptance for publication, an electronic version of a final peer-reviewed, manuscript resulting from research supported in whole or in part, with direct costs from National Institutes of Health. The author's final peer-reviewed manuscript is defined as the final version accepted for journal publication, and includes all modifications from the publishing peer review process. For additional information, please visit <http://publicaccess.nih.gov/>.

Award recipients must promote objectivity in research by establishing standards to ensure that the design, conduct and reporting of research funded under NIH-funded awards are not biased by a conflicting financial interest of an Investigator. Investigator is defined as the Principal Investigator and any other person who is responsible for the design, conduct, or reporting of NIH-funded research or proposed research, including the Investigator's spouse and dependent children. Awardees must have a written administrative process to identify and manage financial conflict of interest and must inform Investigators of the conflict of interest policy and of the Investigators' responsibilities. Prior to expenditure of these awarded funds, the Awardee must report to the NIH Awarding Component the existence of a conflicting interest and within 60 days of any new conflicting interests identified after the initial report. Awardees must comply with these and all other aspects of 42 CFR Part 50, Subpart F. These requirements also apply to subgrantees, contractors, or collaborators engaged by the Awardee under this award. The NIH website <http://grants.nih.gov/grants/policy/coi/index.htm> provides additional information.

If you have any questions about this award, please contact the individual(s) referenced in Section IV.

Sincerely yours,

Victoria Bishton
Grants Management Officer
NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Additional information follows

Award Calculation (U.S. Dollars)

Salaries and Wages	\$840,611
Fringe Benefits	\$237,811
Personnel Costs (Subtotal)	\$1,078,422
Consultant Services	\$83,000
Equipment	\$500,000
Supplies	\$347,000
Travel Costs	\$30,000
Other Costs	\$361,891
Consortium/Contractual Cost	\$553,889

Federal Direct Costs	\$2,954,202
Federal F&A Costs	\$1,318,886
Approved Budget	\$4,273,088
Federal Share	\$4,273,088
TOTAL FEDERAL AWARD AMOUNT	\$4,273,088

AMOUNT OF THIS ACTION (FEDERAL SHARE) \$0

SUMMARY TOTALS FOR ALL YEARS		
YR	THIS AWARD	CUMULATIVE TOTALS
1	\$4,273,088	\$4,273,088
2	\$3,809,825	\$3,809,825
3	\$3,823,170	\$3,823,170
4	\$3,835,522	\$3,835,522
5	\$3,843,821	\$3,843,821

Recommended future year total cost support, subject to the availability of funds and satisfactory progress of the project

Fiscal Information:

CFDA Number: 93.172
 EIN: 1042103580C5
 Document Number: PHG005550A
 Fiscal Year: 2010

IC	CAN	2010	2011	2012	2013	2014
HG	8472563	\$3,878,168	\$3,336,688	\$3,336,512	\$3,335,435	\$3,334,245
HG	8472569	\$394,920	\$473,137	\$486,658	\$500,087	\$509,576

Recommended future year total cost support, subject to the availability of funds and satisfactory progress of the project

NIH Administrative Data:

PCC: X7JS / OC: 414A / Processed eRA Commons
User Name 1/02/2010

SECTION II – PAYMENT/HOTLINE INFORMATION – 1P50HG005550-01 REVISED

For payment and HHS Office of Inspector General Hotline information, see the NIH Home Page at <http://grants.nih.gov/grants/policy/awardconditions.htm>

SECTION III – TERMS AND CONDITIONS – 1P50HG005550-01 REVISED

This award is based on the application submitted to, and as approved by, NIH on the above-titled project and is subject to the terms and conditions incorporated either directly or by reference in the following:

- a. The grant program legislation and program regulation cited in this Notice of Award.
- b. Conditions on activities and expenditure of funds in other statutory requirements, such as those included in appropriations acts.
- c. 45 CFR Part 74 or 45 CFR Part 92 as applicable.

- d. The NIH Grants Policy Statement, including addenda in effect as of the beginning date of the budget period.
- e. This award notice, INCLUDING THE TERMS AND CONDITIONS CITED BELOW.

(See NIH Home Page at '<http://grants.nih.gov/grants/policy/awardconditions.htm>' for certain references cited above.)

This institution is a signatory to the Federal Demonstration Partnership (FDP) Phase V Agreement which requires active institutional participation in new or ongoing FDP demonstrations and pilots.

Carry over of an unobligated balance into the next budget period requires Grants Management Officer prior approval.

In accordance with P.L. 110-161, compliance with the NIH Public Access Policy is now mandatory. For more information, see NOT-OD-08-033 and the Public Access website: <http://publicaccess.nih.gov/>.

This award is funded by the following list of institutes. Any papers published under the auspices of this award must cite the funding support of all institutes.

National Human Genome Research Institute (NHGRI)
--

Treatment of Program Income:

Additional Costs

SECTION IV – HG Special Terms and Conditions – 1P50HG005550-01 REVISED

REVISION #1

The purpose of this revision is to remove the restriction placed on half of the funds awarded for this project (\$2,136,544 total costs) on the award dated 09/12/2010. This award reflects NHGRI receipt and acceptance of the requested information submitted by the grantee via e-mail dated 10/25/2010. Research involving human subjects in this project may begin and consultants may be paid from this award.

THE TERMS AND CONDITIONS STATED BELOW WILL REMAIN IN EFFECT:

1. **RESTRICTION:** Funds awarded for equipment (\$500K direct costs) are restricted and may not be used for any other purpose with the written prior approval of NHGRI.
2. This award is subject to the SPECIAL REQUIREMENTS section of the PAR-08-094, entitled, Centers of Excellence in Genomic Science (CEGS), released on February 22, 2008, which is hereby incorporated by reference as special terms and conditions of this award. Copies of this announcement may be accessed at the following internet address: <http://grants.nih.gov/grants/guide/pa-files/par-08-094.html> or obtained from the Grants Management Contact referenced in the award.
3. Compliance with the data and materials sharing and release plans, described on pages 117-118 and 130-133 of the grant application is a condition of this award. Failure to comply with these plans may result in termination of the award.
4. An annual grantee meeting, to be held at one of the grantee sites or near Washington D.C. is planned for exchange of information among investigators. Awardees will be expected to host these meetings on a rotating basis, as determined in consultation with NIH staff. The principal Investigator is normally expected to attend this meeting; additional center personnel will be invited on a spaceavailable basis. Costs for attending these meetings are included in the grant.
5. The NHGRI is committed to increasing the number of underrepresented minorities trained to conduct genomics and ELSI research as outlined in the Minority Action Plan (MAP; <http://www.genome.gov/10001707>). The proposal submitted by the principal investigator to train NIH underrepresented minorities will be conducted as stated in the approved plan. Funds for this initiative (\$394,920 total costs in the -01 year; \$473,137 total costs -02 year; \$486,658 total costs in the -03 year; \$500,088 total costs in the -04 year; and \$509,576 in the -05 year) are restricted and may not be used for any other purpose without written prior approval of the National Human

Information about the NHGRI's initiative to increase the number of underrepresented minorities trained to conduct genomics and ELSI research and resources to assist investigators in responding to this initiative can be found at: <http://www.genome.gov/10003996>.

6. As part of the annual review of this grant, the principal investigator must submit two reports: a summary report of the MAP activities which is submitted as part of the parent-grant's noncompeting progress report and a more detailed report that is generated for the annual MAP meeting. The principal investigator will be expected to participate in the annual MAP program meeting which is often held in conjunction with the annual meeting of the Centers of Excellence in Genomics Science.

7. Although the budget period start date for this award is September 13, 2010, this award includes funds for 12 months of support. Future year budget periods will cycle on August 1. Grant Progress Reports (PHS 2590) are due two months prior to this date.

8. The facilities and administrative cost portion of the commitments is calculated using the current negotiated Facilities and Administrative cost (F&A) rate(s) dated April 28, 2010.

9. This award includes funds awarded for consortium activity with the Children's Hospital - Boston in the amount of \$80,954 (\$47,089 direct costs + \$33,865 facilities and administrative costs), the University of California San Diego in the amount of \$144,261 (\$93,373 direct costs + \$50,888 facilities and administrative costs), the Massachusetts General Hospital in the amount of \$328,674 (\$192,217 direct costs + \$136,457 facilities and administrative costs). Consortia are to be established and administered as described in the NIH Grants Policy Statement (NIH GPS). The referenced section of the NIH GPS is available at: http://grants.nih.gov/grants/policy/nihgps_2003/NIHGPS_Part12.htm - _Toc54600251

10. In addition to the PI, the following individuals are named as key personnel:

Dr. Kun Zhang
Dr. George Daley
Dr. Keith Joung

Written prior approval is required if any of the individual(s) named above withdraws from the project entirely, is absent from the project during any continuous period of 3 months or more, or reduces time devoted to the project by 25 percent or more from the level that was approved at the time of award.

11. This award is excluded from the Streamlined Non-Competing Award Process.

STAFF CONTACTS

The Grants Management Specialist is responsible for the negotiation, award and administration of this project and for interpretation of Grants Administration policies and provisions. The Program Official is responsible for the scientific, programmatic and technical aspects of this project. These individuals work together in overall project administration. Prior approval requests (signed by an Authorized Organizational Representative) should be submitted in writing to the Grants Management Specialist. Requests may be made via e-mail.

Grants Management Specialist: Lisa A. Oken
Email: loken@mail.nih.gov **Phone:** 301-402-5756 **Fax:** 301-480-1956

Program Official: Jeffery Schloss
Email: schlossj@mail.nih.gov **Phone:** 301-496-7531 **Fax:** 301-480-2770

SPREADSHEET SUMMARY

GRANT NUMBER: 1P50HG005550-01 REVISED

INSTITUTION: HARVARD UNIVERSITY (MEDICAL SCHOOL)

Budget	Year 1	Year 2	Year 3	Year 4	Year 5
Salaries and Wages	\$840,611	\$885,273	\$885,402	\$880,969	\$890,409
Fringe Benefits	\$237,811	\$250,831	\$251,049	\$250,005	\$252,795
Personnel Costs (Subtotal)	\$1,078,422	\$1,136,104	\$1,136,451	\$1,130,974	\$1,143,204
Consultant Services	\$83,000	\$83,000	\$83,000	\$83,000	\$83,000
Equipment	\$500,000				
Supplies	\$347,000	\$364,517	\$373,081	\$380,213	\$371,360
Travel Costs	\$30,000	\$30,300	\$24,609	\$18,927	\$17,255
Other Costs	\$361,891	\$330,002	\$325,835	\$328,655	\$323,146
Consortium/Contractual Cost	\$553,889	\$568,833	\$583,783	\$598,181	\$612,927
TOTAL FEDERAL DC	\$2,954,202	\$2,512,756	\$2,526,759	\$2,539,950	\$2,550,892
TOTAL FEDERAL F&A	\$1,318,886	\$1,297,069	\$1,296,411	\$1,295,572	\$1,292,929
TOTAL COST	\$4,273,088	\$3,809,825	\$3,823,170	\$3,835,522	\$3,843,821

Facilities and Administrative Costs	Year 1	Year 2	Year 3	Year 4	Year 5
F&A Cost Rate ¹	69.5%	69.5%	69.5%	69.5%	69.5%
F&A Cost Base ¹	\$1,897,677	\$1,866,287	\$1,865,340	\$1,864,133	\$1,860,329
F&A Costs ¹	\$1,318,886	\$1,297,069	\$1,296,411	\$1,295,572	\$1,292,929



SPECIALIZED RESEARCH CENTER
Department of Health and Human Services
National Institutes of Health
NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Issue Date: 09/12/2010



Grant Number: 1P50HG005550-01

Principal Investigator(s):
GEORGE M CHURCH, PHD

Project Title: Causal Transcriptional Consequences of Human Genetic Variation

ASSOCIATE DIRECTOR, SPA
HARVARD MEDICAL SCHOOL
25 SHATTUCK STREET
BOSTON, MA 02115

Award e-mailed to: spa_award@hms.harvard.edu

Budget Period: 09/13/2010 – 07/31/2011

Project Period: 09/13/2010 – 07/31/2015

Dear Business Official:

The National Institutes of Health hereby awards a grant in the amount of \$4,273,088 (see "Award Calculation" in Section I and "Terms and Conditions" in Section III) to HARVARD UNIVERSITY (MEDICAL SCHOOL) in support of the above referenced project. This award is pursuant to the authority of 42 USC 241 42 CFR 52 and is subject to the requirements of this statute and regulation and of other referenced, incorporated or attached terms and conditions.

Acceptance of this award including the "Terms and Conditions" is acknowledged by the grantee when funds are drawn down or otherwise obtained from the grant payment system.

Each publication, press release or other document that cites results from NIH grant-supported research must include an acknowledgment of NIH grant support and disclaimer such as "The project described was supported by Award Number P50HG005550 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health."

Award recipients are required to comply with the NIH Public Access Policy. This includes submission to PubMed Central (PMC), upon acceptance for publication, an electronic version of a final peer-reviewed, manuscript resulting from research supported in whole or in part, with direct costs from National Institutes of Health. The author's final peer-reviewed manuscript is defined as the final version accepted for journal publication, and includes all modifications from the publishing peer review process. For additional information, please visit <http://publicaccess.nih.gov/>.

Award recipients must promote objectivity in research by establishing standards to ensure that the design, conduct and reporting of research funded under NIH-funded awards are not biased by a conflicting financial interest of an Investigator. Investigator is defined as the Principal Investigator and any other person who is responsible for the design, conduct, or reporting of NIH-funded research or proposed research, including the Investigator's spouse and dependent children. Awardees must have a written administrative process to identify and manage financial conflict of interest and must inform Investigators of the conflict of interest policy and of the Investigators' responsibilities. Prior to expenditure of these awarded funds, the Awardee must report to the NIH Awarding Component the existence of a conflicting interest and within 60 days of any new conflicting interests identified after the initial report. Awardees must comply with these and all other aspects of 42 CFR Part 50, Subpart F. These requirements also apply to subgrantees, contractors, or collaborators engaged by the Awardee under this award. The NIH website <http://grants.nih.gov/grants/policy/coi/index.htm> provides additional information.

If you have any questions about this award, please contact the individual(s) referenced in Section IV.

Sincerely yours,

Cheryl Chick
Grants Management Officer
NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Additional information follows

Award Calculation (U.S. Dollars)

Salaries and Wages	\$840,611
Fringe Benefits	\$237,811
Personnel Costs (Subtotal)	\$1,078,422
Consultant Services	\$83,000
Equipment	\$500,000
Supplies	\$347,000
Travel Costs	\$30,000
Other Costs	\$361,891
Consortium/Contractual Cost	\$553,889

Federal Direct Costs	\$2,954,202
Federal F&A Costs	\$1,318,886
Approved Budget	\$4,273,088
Federal Share	\$4,273,088
TOTAL FEDERAL AWARD AMOUNT	\$4,273,088

AMOUNT OF THIS ACTION (FEDERAL SHARE) \$4,273,088

SUMMARY TOTALS FOR ALL YEARS		
YR	THIS AWARD	CUMULATIVE TOTALS
1	\$4,273,088	\$4,273,088
2	\$3,809,825	\$3,809,825
3	\$3,823,170	\$3,823,170
4	\$3,835,522	\$3,835,522
5	\$3,843,821	\$3,843,821

Recommended future year total cost support, subject to the availability of funds and satisfactory progress of the project

Fiscal Information:

CFDA Number: 93.172
 EIN: 1042103580C5
 Document Number: PHG005550A
 Fiscal Year: 2010

IC	CAN	2010	2011	2012	2013	2014
HG	8472563	\$3,878,168	\$3,336,688	\$3,336,512	\$3,335,435	\$3,334,245
HG	8472569	\$394,920	\$473,137	\$486,658	\$500,087	\$509,576

Recommended future year total cost support, subject to the availability of funds and satisfactory progress of the project

NIH Administrative Data:

PCC: X7JS / OC: 414A / Processed eRA Commons User Name 09/08/2010

SECTION II – PAYMENT/HOTLINE INFORMATION – 1P50HG005550-01

For payment and HHS Office of Inspector General Hotline information, see the NIH Home Page at <http://grants.nih.gov/grants/policy/awardconditions.htm>

SECTION III – TERMS AND CONDITIONS – 1P50HG005550-01

This award is based on the application submitted to, and as approved by, NIH on the above-titled project and is subject to the terms and conditions incorporated either directly or by reference in the following:

- a. The grant program legislation and program regulation cited in this Notice of Award.
- b. Conditions on activities and expenditure of funds in other statutory requirements, such as those included in appropriations acts.
- c. 45 CFR Part 74 or 45 CFR Part 92 as applicable.

- d. The NIH Grants Policy Statement, including addenda in effect as of the beginning date of the budget period.
- e. This award notice, INCLUDING THE TERMS AND CONDITIONS CITED BELOW.

(See NIH Home Page at '<http://grants.nih.gov/grants/policy/awardconditions.htm>' for certain references cited above.)

This institution is a signatory to the Federal Demonstration Partnership (FDP) Phase V Agreement which requires active institutional participation in new or ongoing FDP demonstrations and pilots.

Carry over of an unobligated balance into the next budget period requires Grants Management Officer prior approval.

In accordance with P.L. 110-161, compliance with the NIH Public Access Policy is now mandatory. For more information, see NOT-OD-08-033 and the Public Access website: <http://publicaccess.nih.gov/>.

This award is funded by the following list of institutes. Any papers published under the auspices of this award must cite the funding support of all institutes.

National Human Genome Research Institute (NHGRI)
--

Treatment of Program Income:

Additional Costs

SECTION IV – HG Special Terms and Conditions – 1P50HG005550-01

1. RESTRICTION: Half of the funds awarded for this project (\$2,136,544 total costs) are restricted pending the submission and NHGRI's review of the following:

IRB letters (either approval or exemption) from both Harvard Medical School and the University of California San Diego;
 Certification of Human Subjects education for key personnel;
 Revised face pages for Harvard Medical School and the University of California San Diego reflecting the correct human subjects status;
 Submission of date of birth for Dr. Dae Kim;
 Submission of consultant rates.

Until the status of human subjects research on this award has been clarified, the grantee institution may conduct only activities that are clearly severable and independent from activities that involve human subjects.

In addition, no funds can be awarded to Dr. Dae Kim until it has been confirmed that he has not been suspended or debarred from federal funding.

Consultants cannot be paid on this award until NHGRI has determined that the rates and work are appropriate and reasonable.

2. RESTRICTION: Funds awarded for equipment (\$500K direct costs) are restricted and may not be used for any other purpose with the written prior approval of NHGRI.

3. This award is subject to the SPECIAL REQUIREMENTS section of the PAR-08-094, entitled, Centers of Excellence in Genomic Science (CEGS), released on February 22, 2008, which is hereby incorporated by reference as special terms and conditions of this award. Copies of this announcement may be accessed at the following internet address: <http://grants.nih.gov/grants/guide/pa-files/par-08-094.html> or obtained from the Grants Management Contact referenced in the award.

4. Compliance with the data and materials sharing and release plans, described on pages 117-118 and 130-133 of the grant application is a condition of this award. Failure to comply with these plans may result in termination of the award.

5. An annual grantee meeting to be held at one of the grantee sites or near Washington, D.C. is planned for exchange of information among investigators. Awardees will be expected to host these meetings on a rotating basis, as determined in consultation with NIH staff. The principal investigator is normally expected to attend this meeting; additional center personnel will be invited on a space-available basis. Costs for attending these meetings are included in the grant.

6. The NHGRI is committed to increasing the number of underrepresented minorities trained to conduct genomics and ELSI research as outlined in the Minority Action Plan (MAP; <http://www.genome.gov/10001707>). The proposal submitted by the principal investigator to train NIH underrepresented minorities will be conducted as stated in the approved plan. Funds for this initiative (\$394,920 total costs in the -01 year; \$473,137 total costs -02 year; \$486,658 total costs in the -03 year; \$500,088 total costs in the -04 year; and \$509,576 in the -05 year) are restricted and may not be used for any other purpose without written prior approval of the National Human Genome Research Institute. Unexpended MAP funds will generally be used to offset future year commitment.

Information about the NHGRI's initiative to increase the number of underrepresented minorities trained to conduct genomics and ELSI research and resources to assist investigators in responding to this initiative can be found at: <http://www.genome.gov/10003996>.

7. As part of the annual review of this grant, the principal investigator must submit two reports: a summary report of the MAP activities which is submitted as part of the parent-grant's noncompeting progress report and a more detailed report that is generated for the annual MAP meeting. The principal investigator will be expected to participate in the annual MAP program meeting which is often held in conjunction with the annual meeting of the Centers of Excellence in Genomics Science.

8. Although the budget period start date for this award is September 13, 2010, this award includes funds for 12 months of support. Future year budget periods will cycle on August 1. Grant Progress Reports (PHS 2590) are due two months prior to this date.

9. The facilities and administrative cost portion of the commitments is calculated using the current negotiated Facilities and Administrative cost (F&A) rate(s) dated April 28, 2010.

10. This award includes funds awarded for consortium activity with the Children's Hospital - Boston in the amount of \$80,954 (\$47,089 direct costs + \$33,865 facilities and administrative costs), the University of California San Diego in the amount of \$144,261 (\$93,373 direct costs + \$50,888 facilities and administrative costs), the Massachusetts General Hospital in the amount of \$328,674 (\$192,217 direct costs + \$136,457 facilities and administrative costs). Consortia are to be established and administered as described in the NIH Grants Policy Statement (NIH GPS). The referenced section of the NIH GPS is available at: http://grants.nih.gov/grants/policy/nihgps_2003/NIHGPS_Part12.htm - _Toc54600251

11. In addition to the PI, the following individuals are named as key personnel:

Dr. Kun Zhang
Dr. George Daley
Dr. Keith Joung

Written prior approval is required if any of the individual(s) named above withdraws from the project entirely, is absent from the project during any continuous period of 3 months or more, or reduces time devoted to the project by 25 percent or more from the level that was approved at the time of award.

12. This award is excluded from the Streamlined Non-Competing Award Process.

STAFF CONTACTS

The Grants Management Specialist is responsible for the negotiation, award and administration of this project and for interpretation of Grants Administration policies and provisions. The Program Official is responsible for the scientific, programmatic and technical aspects of this project. These individuals work together in overall project administration. Prior approval requests (signed by an Authorized Organizational Representative) should be submitted in writing to the Grants Management Specialist. Requests may be made via e-mail.

Grants Management Specialist, Debbie P. Chen
 Email: chendeb@mail.nih.gov Phone: 301-594-5250 Fax: 301-451-5434

Program Official: Jeffery Schloss
 Email: schlossj@mail.nih.gov Phone: 301-496-7531 Fax: 301-480-2770

SPREADSHEET SUMMARY
GRANT NUMBER: 1P50HG005550-01

INSTITUTION: HARVARD UNIVERSITY (MEDICAL SCHOOL)

Budget	Year 1	Year 2	Year 3	Year 4	Year 5
Salaries and Wages	\$840,611	\$885,273	\$885,402	\$880,969	\$890,409
Fringe Benefits	\$237,811	\$250,831	\$251,049	\$250,005	\$252,795
Personnel Costs (Subtotal)	\$1,078,422	\$1,136,104	\$1,136,451	\$1,130,974	\$1,143,204
Consultant Services	\$83,000	\$83,000	\$83,000	\$83,000	\$83,000
Equipment	\$500,000				
Supplies	\$347,000	\$364,517	\$373,081	\$380,213	\$371,360
Travel Costs	\$30,000	\$30,300	\$24,609	\$18,927	\$17,255
Other Costs	\$361,891	\$330,002	\$325,835	\$328,655	\$323,146
Consortium/Contractual Cost	\$553,889	\$568,833	\$583,783	\$598,181	\$612,927
TOTAL FEDERAL DC	\$2,954,202	\$2,512,756	\$2,526,759	\$2,539,950	\$2,550,892
TOTAL FEDERAL F&A	\$1,318,886	\$1,297,069	\$1,296,411	\$1,295,572	\$1,292,929
TOTAL COST	\$4,273,088	\$3,809,825	\$3,823,170	\$3,835,522	\$3,843,821

Facilities and Administrative Costs	Year 1	Year 2	Year 3	Year 4	Year 5
F&A Cost Rate 1	69.5%	69.5%	69.5%	69.5%	69.5%
F&A Cost Base 1	\$1,897,677	\$1,866,287	\$1,865,340	\$1,864,133	\$1,860,329
F&A Costs 1	\$1,318,886	\$1,297,069	\$1,296,411	\$1,295,572	\$1,292,929

We use this update to our proposal for a Center for the Causal Transcriptional Consequences of Human Genetic Variation (CTCHGV) to describe new preliminary results relevant to our four Aims. We summarize these here and follow up with more detailed descriptions.

Aim 1: (A) Using samples and genome sequence from the Personal Genome Project (PGP), we are performing a “trial run” of Aim 1 procedures for identifying variations *cis* to genes exhibiting allele-specific expression (ASE) and using Zinc Finger Nucleases (ZFNs) to alter the variations and assess their role in actually causing ASE. This trial run is giving us experience in performing Aim 1 methods and helping to define computational procedures. (B) We have generated 13 additional TNN and ANN OPEN zinc finger pools and will shortly obtain CHIP-Seq data that will yield comprehensive information on locations of off-target ZFN sites. (C) We report improved protocols for quantifying ASE and analysis of ASE in HapMap populations. (D) We are experimenting with genotyping of single cells via *in situ* Rolling Circle Amplification (RCA) and MIP-based detection of RNA alleles from RNA vs. cDNA templates. (E) Our Multiplex Automated Genome Engineering (MAGE) method was published (11) and called “a major advance in synthetic biology” (9). In MAGE, multiple targeted alterations are simultaneously and efficiently made in the *E. coli* genome by transfecting pools of oligos that are synthesized as copies of the targeted regions that include the desired modifications. We have proposed using MAGE on CTCHGV subject BACS in *E. coli* to generate combinatorially altered *cis* gene regions that are then recombined back into subject cell line genomes using ZFNs. Here we report initial efforts to apply MAGE directly to human cells, which we proposed as a second method for altering *cis* gene regions.

Aim 2: We have developed a prototype of a set of flow cell chambers capable of immortalizing and reprogramming a population panel of human skin fibroblasts into iPS.

Aim 3: We are experimenting with split pool synthesis of oligonucleotides that we expect will provide the basis for generating the bar-coded sequencing primer sets required by the first of our three Aim 3 approaches to single cell transcriptomics. Note that Aim 1 (D) above also serves as a preliminary step towards the second of our Aim 3 approaches, which focuses on single cell *in situ* RCA amplification of transcripts.

Aim 4: As initial steps towards the integrated synthesis and sequencing platform, we are in the process of outfitting a Polonator with a Digital Micromirror Device and experimenting with photo-labile sequence bead surface attachment chemistries.

Detailed descriptions of these new preliminary results

Aim 1: A. Trial run of Aim 1 procedures: Using ASE data described in section 4.3 of our proposal we identified 34 genes in subject PGP1 that consistently exhibit ASE in twenty fibroblast, fibroblast-derived iPS, and variously differentiated iPS cell lines. Meanwhile, a complete genome sequence for PGP1 was obtained as a donation from Complete Genomics Inc. (CGI). We are now using these resources as the basis of our trial run of Aim 1 procedures.

Quality of the CGI PGP1 genome: Although CGI is developing technology for delivery of phased diploid genome sequences, the PGP1 genome is not phased. We have analyzed CGI-delivered summary files of variations and sequence-called regions provided by CGI. Of non-autosomal bases called in the reference genome, 90.7% were called in the CGI genome. In addition to explicitly specified variations, CGI describes 134.1Mb (~4.7%) of the genome incompletely as merely “consistent” or “inconsistent” with the reference sequence. Comparisons with 500K Affy SNP Chip data on PGP1 indicates > 99.3% concordance in all regions fully called by CGI, and > 93.5% if uncalled or incompletely described regions are counted as false negatives.

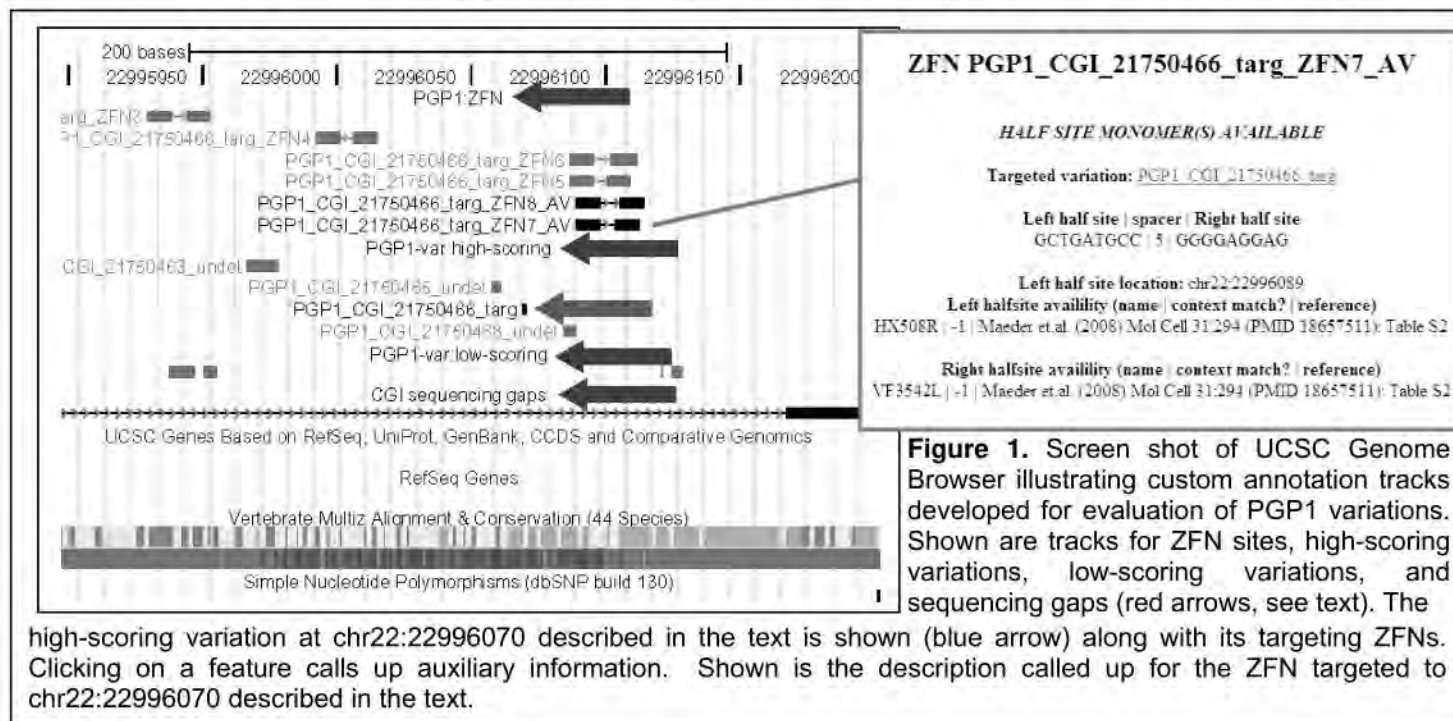
Computational identification of variations potentially causing ASE: We consolidated CGI data on PGP1 sequence variations with gene region annotations, “phyloP” conservation data (10) obtained from alignments of vertebrate genomes, and Transfac-based transcription factor binding site predictions downloaded from the UCSC Genome Browser (3, 4), focusing on the 34 ASE-exhibiting genes. We plan to integrate CHIP² data from the Genome Browser as well as *cis*-eQTL data encompassing data sources from <http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/> and (1). We then identified 141 CGI-identified variations in the regions extending from 100Kbp upstream to 100Kbp downstream of these genes that were “high scoring” by virtue of either being highly conserved or in a Transfac binding site, where “highly conserved” was interpreted as having a phyloP score $\geq 40^{\text{th}}$ percentile of phyloP scores for all codon position 1 and 2 bases in the 34

Program Director/Principal Investigator (Last, First, Middle): Church, George M. CEGS Proposal Update Sept.24,2009

coding regions. CGI sequence calls covered between 91.3% (introns) and 95.3% (exons) of the 34 genes, with the 100Kbps up- and down-stream regions covered at intermediate levels (92.2% and 93.6%, respectively), so this procedure was close to comprehensive in its consideration of all variants in or near *cis*- to these genes.

Zinc Finger Nuclease (ZFN) target site identification; initial selection of variations to alter: While we propose high-throughput methods to develop, optimize, and synthesize ZFNs for application to altering potentially causative variations, for purposes of the trial run we simply identified sites up to 500bp away from our 141 variations that could be targeted by Zinc Finger Consortium "OPEN" pool ZFN arrays (6). Of the 141 variations, 102 could be so targeted. One variation at chr22:22996070 was 31bp away from a site comprising halfsites HX508R and VF3542L published in (6) separated by 5 bp for which ZFN monomers have already been generated and tested by the Joung Lab. CGI identifies the variation as a heterozygous T→C SNP that is not in dbSNP. It is ~715bp upstream of PGP1 ASE exhibiting gene NM_015330, has a phyloP score that is at the 48th percentile of codon 1 and 2 coding bases, but is not in any Transfac-predicted binding sites. We are presently validating the PGP1 CGI sequence at and near chr22:22996070 and are obtaining plasmids for the ZFN monomers. We plan to then introduce and express the ZFNs and verify cutting at the target site by limited cycle PCR around the target, followed by cloning in *E. coli* and sequencing to identify sequence variations attributable to non-homologous end-joining (NHEJ), procedures used in (2, 6). Following verification of cutting, we will clone both PGP1 haplotypes around the variation and use the ZFN to make cell lines homozygous in each of the alleles using the single cell isolation and culturing techniques outlined in Aim 1, and then test for effects on ASE. If this procedure is successful, we will follow up by repeating them at one or more of the other 140 PGP1 variations and will ourselves generate one or both ZFN monomers using the OPEN procedures.

Analysis tool: To help guide selection of subsequent targets for alteration, we have developed custom annotation tracks for the UCSC Genome Browser that show variations, CGI sequence coverage gaps, and ZFN sites for our 34 ASE exhibiting genes. The neighborhood of our target variation can be seen in Figure 1.



B. ZFN development: Since submission of our proposal, we have generated 13 new zinc finger OPEN pools for TNN and ANN subsites (see Figure 2). Also, in addition to the CHIP-Seq experiment described in section 4.5 of our proposal for identification and quantification of target vs. non-target binding for a VEGF-A targeting ZFN, we have now conducted a similar experiment using a pair of four-finger ZFNs made by Sangamo BioSciences targeted to the CCR5 gene. Initial qPCR results showed a 42-fold enrichment of the CCR5 target site in the on the CHIPed DNA compared with whole cell extract (comparable to 36-fold enrichment for VEGF-A reported in

our proposal. For both experiments, CHIPed DNA was delivered to the Broad Institute for Illumina sequencing and we expect to begin analysis shortly.

C. Additional ASE results:

Molecular Inversion Probe

(MIP) improvements:

The MIP gap defines the region in which template DNA is copied from the regions targeted by MIP arms. We measure ASE for thousands of genes at

once by designing MIPs targeted to regions containing heterozygous cDNA SNPs and comparing the read counts containing the two alleles after targeted sequencing using these MIPs (13). We have found that bias can result if the SNPs are too close to the ligation terminus. Figure 3 presents recent data showing that bias is considerably reduced by widening the MIP gap to 18 bases. Unpublished

Unpublished

samples: In collaboration with T. Pastinen (McGill University) and B. Stranger (Harvard) we are identifying genes and regulatory variants associated with ASE in HapMap samples that are prevalent in particular populations or cell types. We are also examining gene expression networks in these HapMap samples to find *cis*-regulatory loci most likely to produce a significant expression perturbation downstream. So far, we found that the 34 ASE exhibiting genes identified from PGP1 also exhibit strong associations with ASE in HapMap lymphocyte samples, and 27 of the 32 were replicated in a *cis*-eQTL study using eight different HapMap populations.

D. Single cell *in situ* genotyping and isolation of cells bearing specific genotypes; RNA allele detection: We are performing proof of concept experiments in single cell *in situ* genotyping and RNA allele detection that are applicable to the assaying of combinatorially modified cell populations described in Aim 1.2 and the single cell isolation and culturing described in Aim 1.1.

Single cell genotyping and cell isolation: We mixed cell lines in 1:1000 proportions derived from two individuals, one of whom is homozygous and the other heterozygous for SNPs rs7825439, and added and centrifuged $\sim 10^4$ total cells to

F1				F2				F3			
GAA	GCA	GGA	GTA	GAA	GCA	GGA	GTA	GAA	GCA	GGA	GTA
GAC	GCC	GGC	GTC	GAC	GCC	GGC	GTC	GAC	GCC	GGC	GTC
GAG	GCG	GGG	GTG	GAG	GCG	GGG	GTG	GAG	GCG	GGG	GTG
GAT	GCT	GGT	GTT	GAT	GCT	GGT	GTT	GAT	GCT	GGT	GTT
TAA	TCA	TGA	TTA	TAA	TCA	TGA	TTA	TAA	TCA	TGA	TTA
TAC	TCC	TGC	TTG	TAC	TCC	TGC	TTG	TAC	TCC	TGC	TTG
TAG	TCE	TGG	TTG	TAG	TCG	TGG	TTG	TAG	TCG	TGG	TTG
TAT	TCT	TGT	TTT	TAT	TCT	TGT	TTT	TAT	TCT	TGT	TTT
AAA	ACA	AGA	ATA	AAA	ACA	AGA	ATA	AAA	ACA	AGA	ATA
AAC	ACC	AGC	ATC	AAC	ACC	AGC	ATC	AAC	ACC	AGC	ATC
AAG	ACG	AGG	ATG	AAG	ACG	AGG	ATG	AAG	ACG	AGG	ATG
AAT	ACT	AGT	ATT	AAT	ACT	AGT	ATT	AAT	ACT	AGT	ATT

Figure 2: New OPEN zinc finger pools for TNN and ANN subsites (red). Pools generated in (6) are shown in gray.

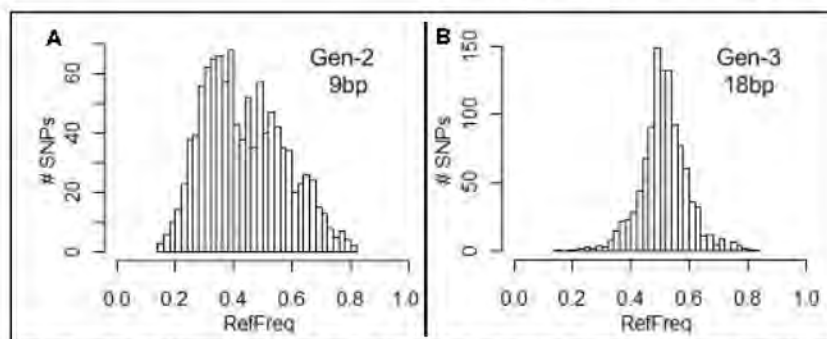


Figure 3: ASE quantification bias is reduced by widening the MIP gap defining the region in which template cDNA containing SNPs is copied. By increasing the 9 bp gap of our "Gen 2" MIPs in (13) (A) to 18 bp (B), so that the target SNPs are outside the 12bp ligase footprint, the distribution of ASE values more closely follows a Gaussian distribution with mean 0.5.

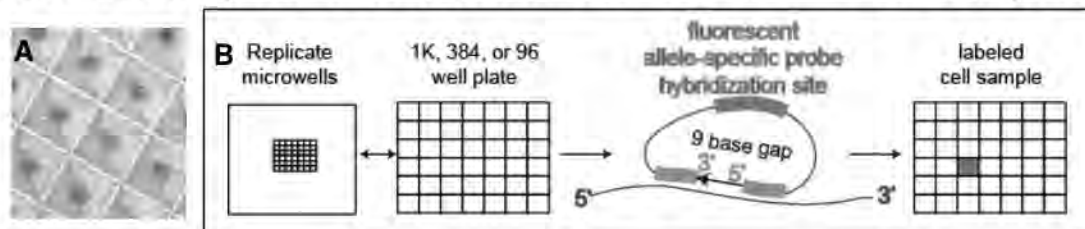


Figure 4: Screening and isolation of mammalian cells with natural and synthetic allelic variants in the genomic DNA. (A) Image of Aggrewell 400 microwells (StemCells, Inc.) plates used to split 50,000 cells into 1,200 microwells. (B) Process for analyzing cells in Aggrewell microwells for genomic variants and ASE. Allele-specific MIPs are circularized against RNA or DNA of Aggrewell microwell contents transferred from replicates to 96 well plates, and amplicons are assayed via MIP-specific probes.

Aggrewell 400 microwells (400- μ M; StemCell Technologies) (~10 cells per well, 1,200 wells; see Figure 4A). The cells were allowed to proliferate and grow onto a tissue culture cover slip, creating a 'replicate' cellular array; the cells on the cover slip were then fixed and the microwells covered with a new cover slip. We then heated the array of fixed cells to 95°C, added allele-specific MIPs for rs7825439, performed normal MIP hybridization, extension, and ligation procedures to create circularized MIPs copying template rs7825439 DNA, and digested remaining uncircularized probes *via* exonuclease. We then conducted probe-specific RCA *in situ* and hybridized allele-specific fluorescent primers to detect cells with heterozygous 'changed' alleles in the microwell replicates. Initial results indicated inefficient circle formation but we expect to overcome this using the method of (5). When optimized, these techniques will support the method proposed in Aim 1.2 and have potential to enable isolation of single cells with distinct genotypes *via* photoablation from replicate cover slips or microplates as required by Aim 1.1. (See Figure 4B.) *MIP-based RNA-allele detection*: We have also shown that MIPs are able to directly circularize using RNA molecules as template and reverse transcriptases (see Figure 5). These methods are required in connection with procedures outlined in section 5.1.2 of our proposal. They will also generally enable targeted assays of non-coding RNAs in addition to mRNAs.

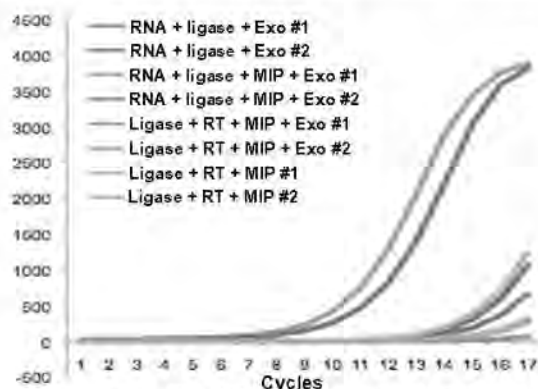


Figure 5: RT-PCR results demonstrating specificity and the sensitivity of detection of MIPs from RNA templates. Performance is comparable to MIP-based detection of genomic DNA alleles.

E. Initial development of MAGE in human cells: We have conducted experiments whose results confirm previous reports that NHEJ (8) and mismatch repair (7) pathways negatively affect oligo-mediated gene targeting. We transfected HeLa cells with a stably integrated mutated EGFP (12) with shRNAs (two each) targeting Ku70, Ku86, and MSH2. After 48 hours, we transfected the cells with oligos targeting the non-transcribed strand of the mutated EGFP containing a mismatch restoring EGFP function. After another 48 hours, we measured repair efficiency in 50,000 cells using flow cytometry, excluding non-viable cells by propidium iodide staining and normalizing data relative to the background fluorescence of untreated cells. The results (see Table 1) support our proposed strategy of using knockdowns / knockouts of competing repair pathways to increase efficiency of oligo-mediated recombination. We are now exploring whether synergy is achieved by targeting both repair pathways simultaneously, and if overexpressing Rad51 has an analogous effect to the β protein of our *E. coli* MAGE system.

RNAi		
Target	shRNA1	shRNA2
Ku86	0.37 \pm 1.38E-02	0.10 \pm 2.59E-02
Ku70	0.47 \pm 5.00E-02	0.62 \pm 3.44E-01
MSH2	0.43 \pm 6.86E-02	0.38 \pm 1.82E-04
Controls		
No oligo / no shRNA	0.00 \pm 3.60E-02	
Oligo / no shRNA	0.19 \pm 1.31E-01	

Table 1: Knockdown of NHEJ and mismatch repair pathway proteins increases efficiency of oligo-mediated recombination in HeLa cells. Values represent percentages of cells repaired. Standard deviations based on $n=2$ replicates.

Aim 2: We have been developing prototypes for the automated iPS management system described in Aim 2 of our proposal. Figure 6 shows a prototype automated fluidics platform for propagating and distributing panels of human skin fibroblasts in which microspheres coated with bovine collagen or Matrigel are used to maintain individual cell samples. Our plans provide for use of fluidics to enable automated handling of the cells, including electroporation- and lentivirus-based iPS transformation, while the transparent sample tubes will enable microscopic and optical monitoring of cell growth.

Aim 3: Split pool synthesis for single cell transcriptomics primer barcoding: Our strategy for split pooled DNA synthesis on bead surfaces integrates three atypical oligonucleotide synthesis strategies: (i) split pooled nucleotide additions, (ii) synthesis on a solid-support surface rather than controlled pore glass (CPG), and (iii) 5'-3' oligonucleotide synthesis. We have begun developing the first two of these protocols: (i) Split pooled synthesis of a 76mer containing 12 split pooled additions was successful (see Figure 7). Following reverse-phase purification to select for only full-length oligonucleotides containing a DMT group, the yield was approximately 320 nmoles, consistent with the expected yield for an oligonucleotide of this size. Trityl yield

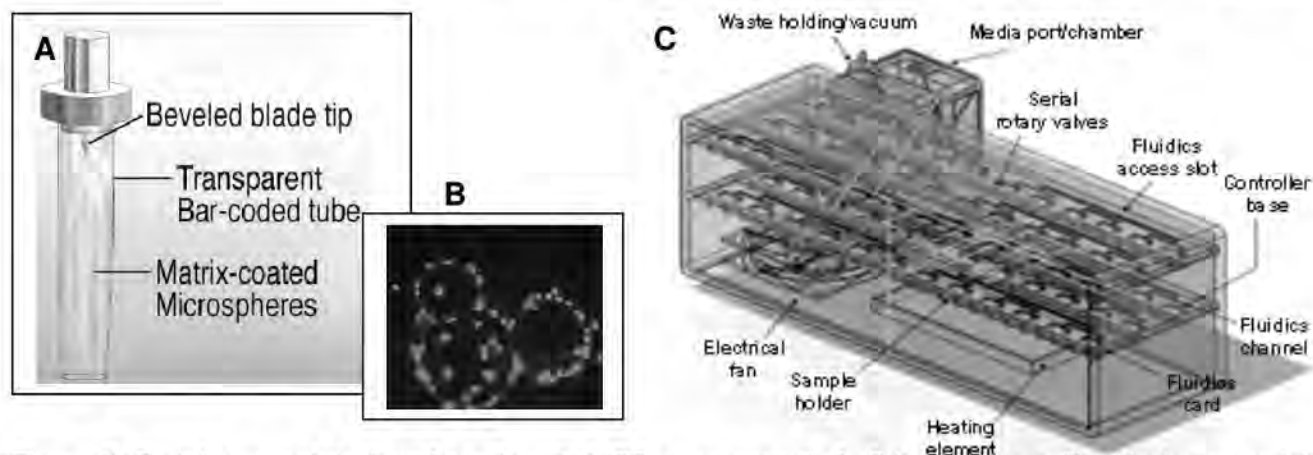
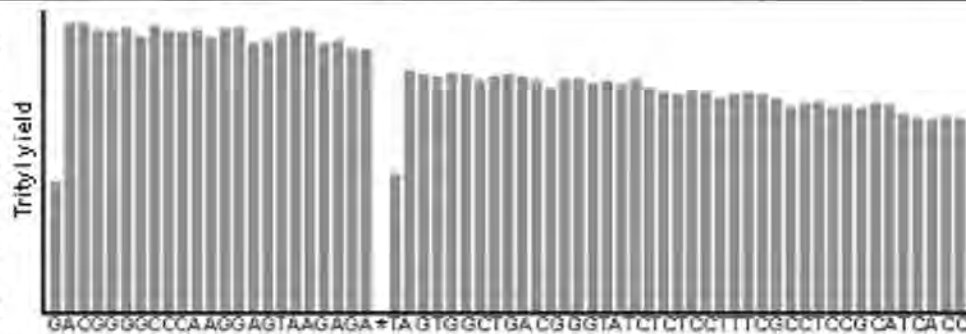


Figure 6. Prototypes and designs for automated iPS management. **A.** Primary human fibroblasts are maintained on microspheres compatible with liquid handling devices. Fibroblasts remain viable at room temperature for up to one week. **B.** Fibroblasts (DAPI stained) growing on microspheres. **C.** Design for automated cell derivation, propagation, maintenance and limited manipulation of primary human fibroblasts. Tissue samples maintained using the device in A will be directly connected to the heated fluidics handler, which will automate media exchanges, cell-microsphere sampling for growth, and the addition of different types of microspheres (e.g., tagged, magnetic). The current prototype will process up to 144 samples independently, using a series of four 3-way rotary valves. This module will then be linked to another fluidics module capable of imaging both adherent and suspended cells in a flow cell during iPS reprogramming. The prototyped model is shown above without attached motors.

Figure 7: Trityl yield during synthesis of 76mer including 12 split pooled additions. Following the first 23 additions, the synthesis was paused and the column removed from the synthesizer for 12 rounds of split pooled synthesis, after which synthesis was then resumed on the synthesizer. Trityl yields (arbitrary units) are shown for additions prior to (left) and after (right, red bar) the split pool additions; an asterisk (*) indicates the split pool additions. Low yield measurements for initial additions in the synthesizer are a common artifact caused by the dryness of the column.



showed minimal inefficiencies. We are currently analyzing the resulting population of molecules *via* multiplexed DNA sequencing, using the fixed sequences synthesized on either side of the split pool additions as primers. (ii) In addition, a synthesis protocol for on-surface synthesis has been developed, requiring extensive washing between synthesis steps as the polystyrene bead substrate was not optimal with standard protocols. Successful synthesis was analyzed by the hybridization of a fluorescently labeled complementary probe (data not shown). We are currently examining split pooled synthesis on such substrates.

Aim 4: Integrated sequencing / synthesis: To assess the feasibility of the integrated sequencing / synthesis approach proposed in Aim 4.1, we have designed, manufactured, and assembled custom optical hardware that incorporates a Digital Micro-mirror Device (DMD) (TI 1080p UV-version) into a Polonator as described in Figure 6.4.1-2 of our proposal and have been writing required support software: low-level Linux drivers, an image generator that accepts bead/colony coordinates, and functions to compute DMD coordinates from image data. Additional components to be written will support automatic alignment of DMD and camera coordinates, network communication between the processor and acquisition to control photorelease, and routines to choose beads for release. Initial testing indicates that the DMD is now sufficiently focused that an image selecting 1 μ m bead-sized regions can be projected onto the Polonator's imaging plane. We are also experimenting with different attachment chemistries (including hetero-bifunctional linkers, homo-bifunctional linkers, and click chemistry) and protocols to develop simple, efficient methods to tether polony microbeads to the Polonator array by photo-labile linkers.

REFERENCES

1. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325: 1246-50.
2. Foley JE, Yeh JR, Maeder ML, Reyon D, Sander JD, Peterson RT, Joung JK. 2009. Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN). *PLoS ONE* 4: e4348. PMC ID: PMC2634973.
3. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* 12: 996-1006. PMC ID: PMC186604.
4. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* 37: D755-61.
5. Larsson C, Koch J, Nygren A, Janssen G, Raap AK, Landegren U, Nilsson M. 2004. In situ genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. *Nat Methods* 1: 227-32.
6. Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichinger M, Jiang T, Foley JE, Winfrey RJ, Townsend JA, Unger-Wallace E, Sander JD, Muller-Lerch F, Fu F, Pearlberg J, Gobel C, Dassie JP, Pruett-Miller SM, Porteus MH, Sgroi DC, Iafrate AJ, Dobbs D, McCray PB, Jr., Cathomen T, Voytas DF, Joung JK. 2008. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* 31: 294-301. PMC ID: PMC2535758.
7. Maguire KK, Kmiec EB. 2007. Multiple roles for MSH2 in the repair of a deletion mutation directed by modified single-stranded oligonucleotides. *Gene* 386: 107-14. PMC ID: PMC1847641.
8. Morozov V, Wawrousek EF. 2008. Single-strand DNA-mediated targeted mutagenesis of genomic DNA in early mouse embryos is stimulated by Rad51/54 and by Ku70/86 inhibition. *Gene Ther* 15: 468-72.
9. Pennisi E. 2009. Genetic engineering. Two steps forward for synthetic biology. *Science* 325: 928-9.
10. Siepel A, Pollard KS, Haussler D. 2006. *New methods for detecting lineage-specific selection* (<http://compgen.bscb.cornell.edu/~acs/dless.pdf>). Presented at Tenth International Conference on Research in Computational Molecular Biology (RECOMB 2006), Venice, Italy
11. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM. 2009. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460: 894-8.
12. Yin WX, Wu XS, Liu G, Li ZH, Watt RM, Huang JD, Liu DP, Liang CC. 2005. Targeted correction of a chromosomal point mutation by modified single-stranded oligonucleotides in a GFP recovery system. *Biochem Biophys Res Commun* 334: 1032-41.
13. Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, Leproust EM, Eggan K, Church GM. 2009. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* 6: 613-8.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

Proposal for a Center for the determination of the
Causal Transcriptional Consequences of Human Genetic Variation (CTCHGV)

submitted
May 25, 2009
by

Professor George M. Church
Department of Genetics
Harvard Medical School

in response to Program Announcement Number
PAR-08-094

COLOR FIGURES

Program Director/Principal Investigator (Last, First, Middle: Church, George M.

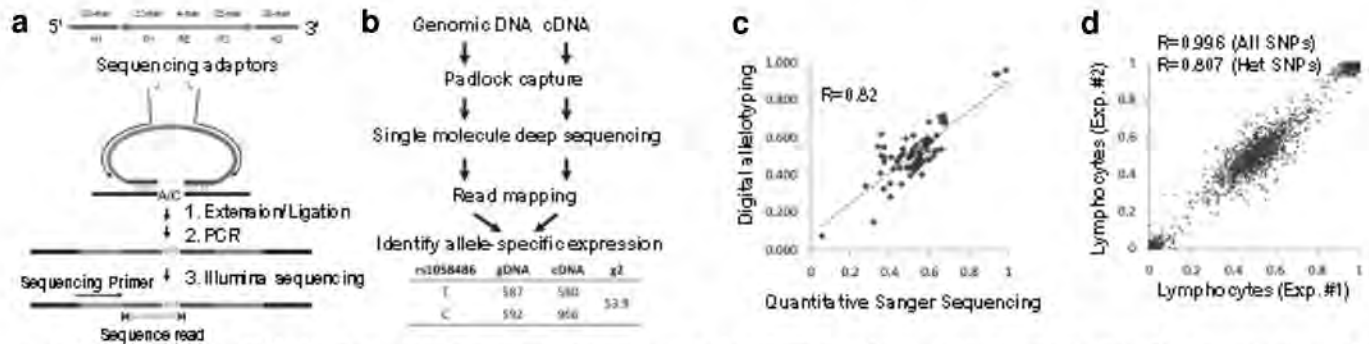


Figure 4.2-1. RNA allelotyping. (a) A schematic diagram for MIP capture and single-molecule sequencing. (b) Detection of allele-specific gene expression. (c) Comparison of allelic ratios measured by RNA allelotyping and quantitative Sanger sequencing. (d) Comparison of allelic ratios between technical replicates.

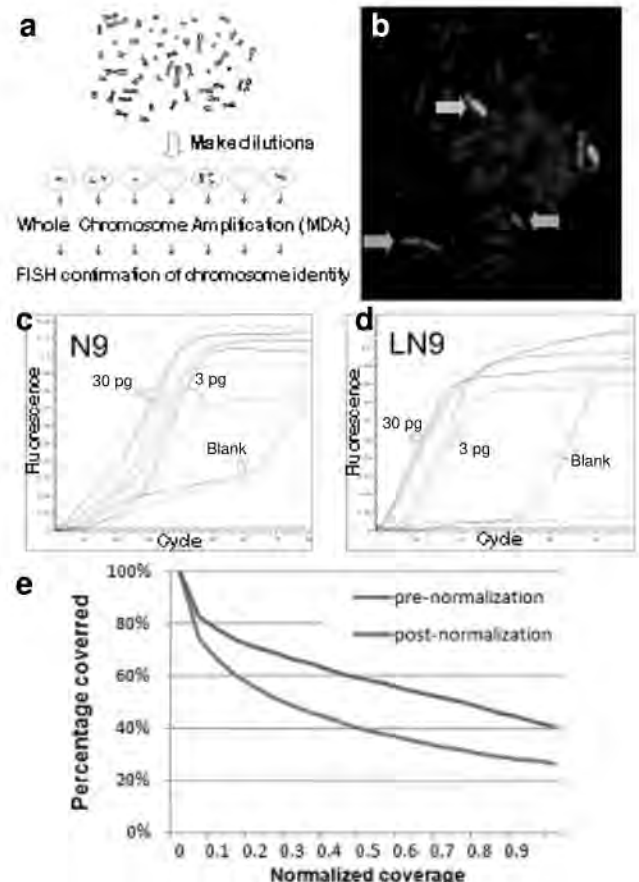


Figure 4.2-2. Polymerase cloning of human chromosome molecules. (a) MDA is performed on limited dilutions of human metaphase chromosome molecules. (b) FISH hybridization confirmed that one amplicon (purple) was from chromosome 6 and another amplicon (green) was from chromosome 19. (c, d) Real-time amplification curves with the N9 and LN9 primers. (e) Reduction of representation bias with post-amplification normalization.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

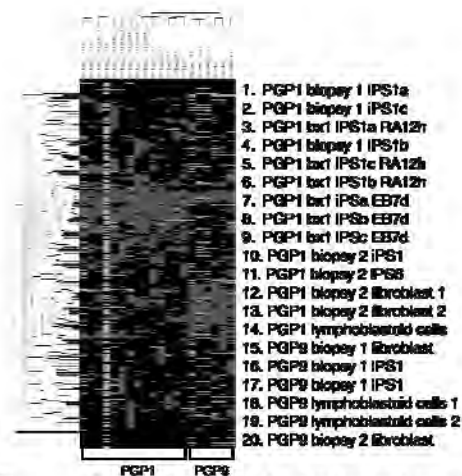


Figure 4.3-1. Hierarchical clustering of statistically significant allele-specific expression (ASE) in reprogrammed cells, showing that ~50% of overall ASE signature was invariant among different cell types, culture conditions and cell batches.

Program Director/Principal Investigator (Last, First, Middle: Church, George M.

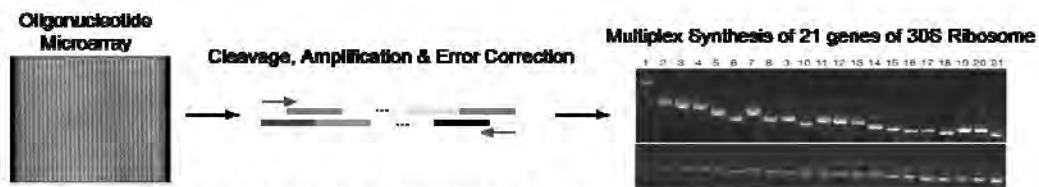


Figure 4.4.1-1 Multiplex DNA Synthesis or large DNA fragments (174)

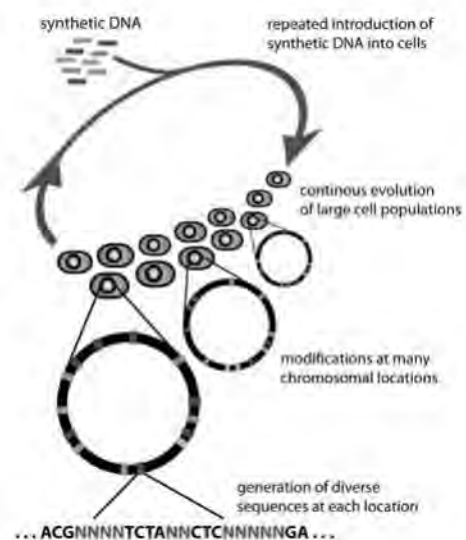


Figure 4.4.2-1 Multiplex Automated Genome Engineering (MAGE)

Program Director/Principal Investigator (Last, First, Middle: Church, George M.

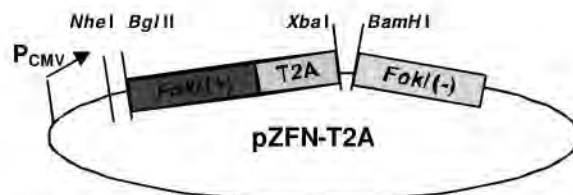


Figure 4.5-1. pZFN-T2A – dual ZFN expression vector

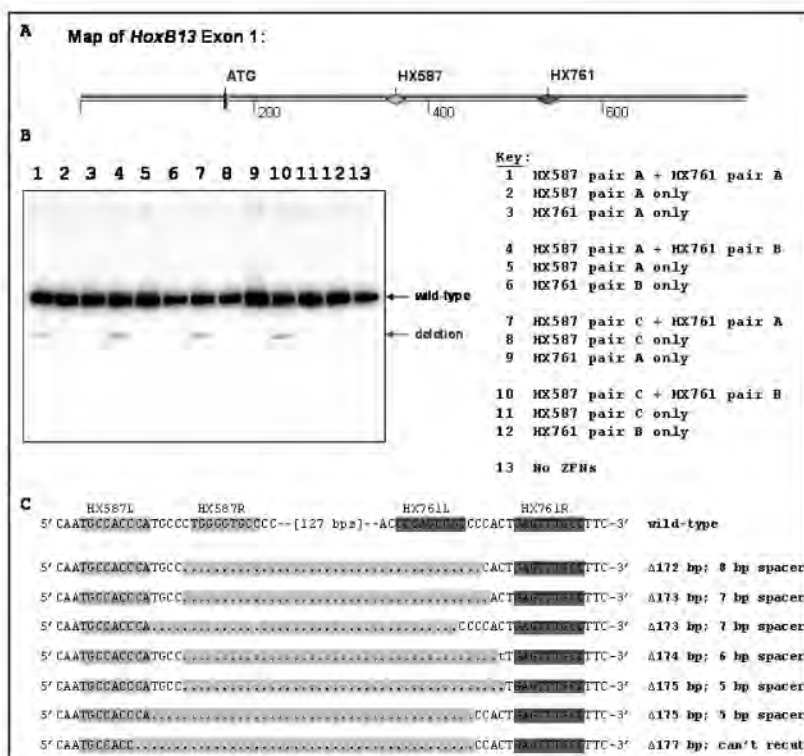


Figure 4.5-2. Dual cleavage of an endogenous *HoxB13* allele leads to deletion of intervening sequence. **A.** Map of human *HoxB13* exon 1 with sites targeted by ZFNs. **B.** Limited-cycle PCR assay of genomic DNA from cells treated with ZFN combinations of ZFNs that cut at the HX587 and HX761 sites (107). **C.** DNA sequences of deletion alleles cloned from genomic DNA of cells treated with two ZFN pairs that cleave at the HX587 (blue) and HX761 (pink) sites. ZFN half-sites are highlighted for each full ZFN site: left (L), right (R). Deletions indicated in grey.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

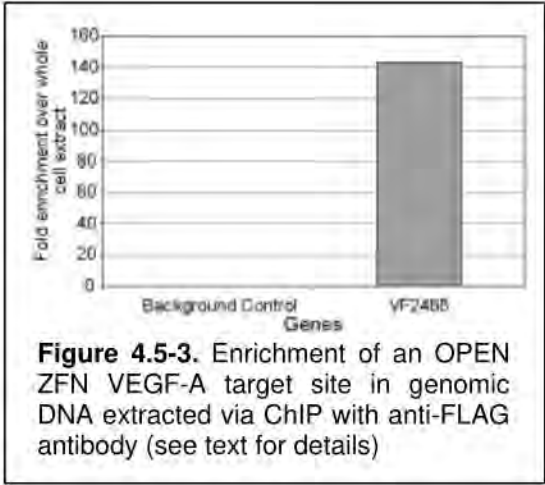


Figure 4.5-3. Enrichment of an OPEN ZFN VEGF-A target site in genomic DNA extracted via ChIP with anti-FLAG antibody (see text for details)

Program Director/Principal Investigator (Last, First, Middle: Church, George M.

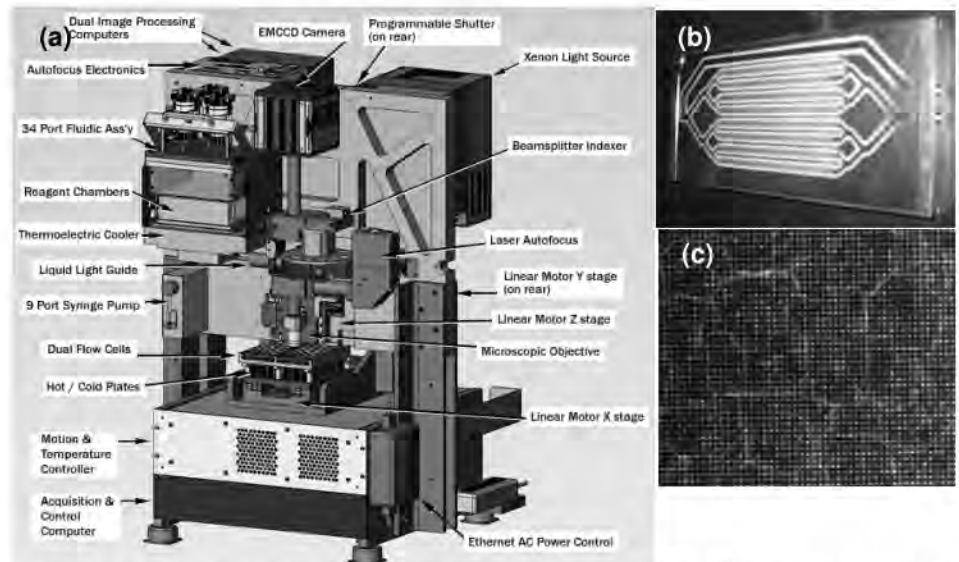


Figure 4.6-1. (a) Polonator instrument. **(b)** Flow cell designed and incorporated into Polonator to increase throughput and reduce runs costs by reducing reagent volumes and expenses. Overall dimensions: 150 X 70 X 8 (mm); lanes' "active area": 70 X 3.3 X 0.1 (mm) (.025 mm in testing). When loaded, the flow cell contains 0.5-1e9 1 μ m beads. **(c)** Polonies created by Rolling Circle Amplification (RCA) instead of on 1 μ m beads, deposited on a grid with 600 nm spot diameter and center to center spacing of 1700 nm. The grid was etched on a silicon wafer using standard photolithography. The image was obtained after a single Sequencing by Synthesis cycle on the Polonator using fluorescent reversible terminators (112, 190). Different colors represent the different bases incorporated during the cycle. Note that CTCHGV proposes to develop single cell transcriptomics using RCA polonies in Aim 3.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

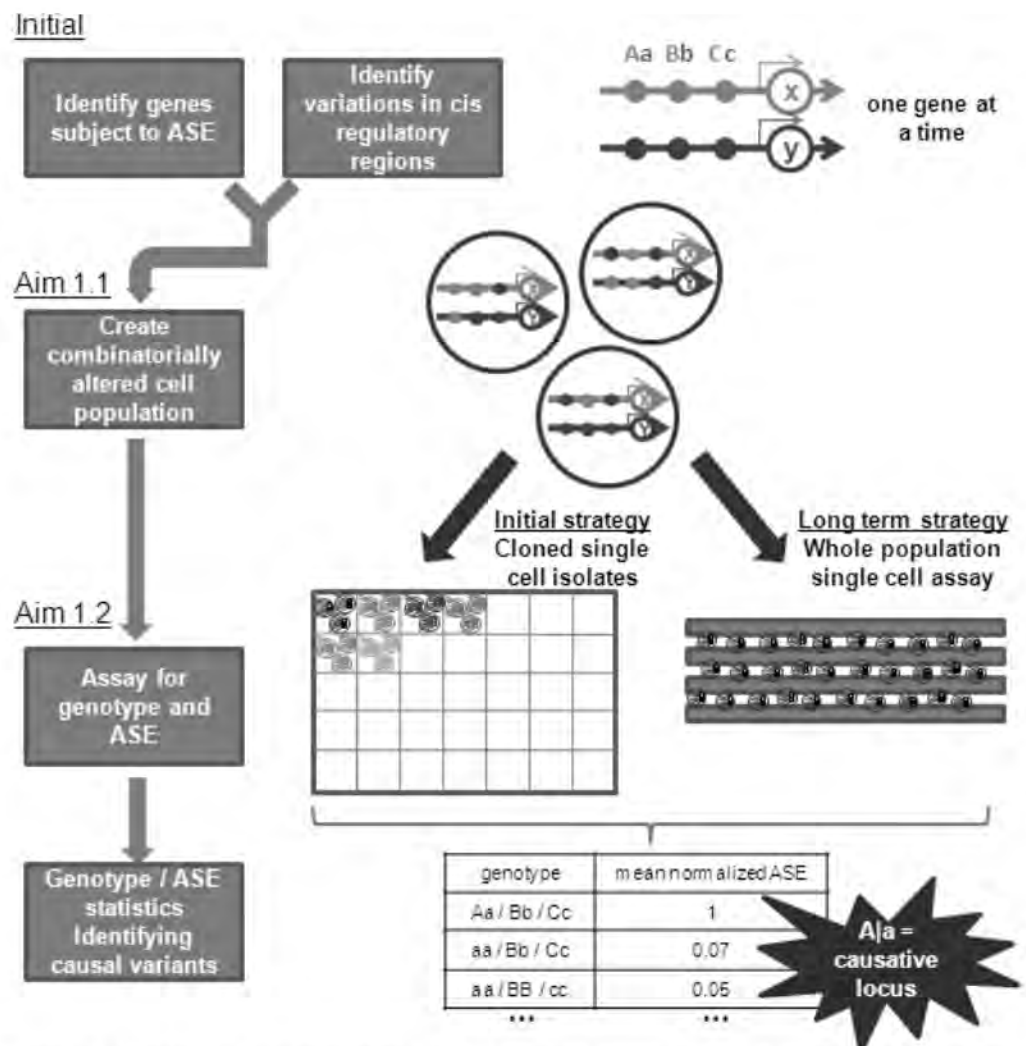


Figure 5-1: Overview of Aim 1 strategy for identifying causative cis variations. Initial Aim 1 work identifies genes subject to allele-specific expression (ASE), and, via next-gen sequence data, also identifies variations in regulatory regions, e.g., here the 100kpbs upstream region. Via Aim 1.1, cell populations are created for each gene bearing combinations of the variations identified for that gene. Via Aim 1.2, cells from this population are genotyped and assessed for ASE so that the specific loci and loci interactions that control ASE can be identified. The initial Aim 1.2 strategy will examine clonal outgrowths of individual altered cells from the population, while a longer term strategy will assay the entire mixed altered population at a single cell level. This strategy is executed one gene at a time for 100s to 1000s of genes.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

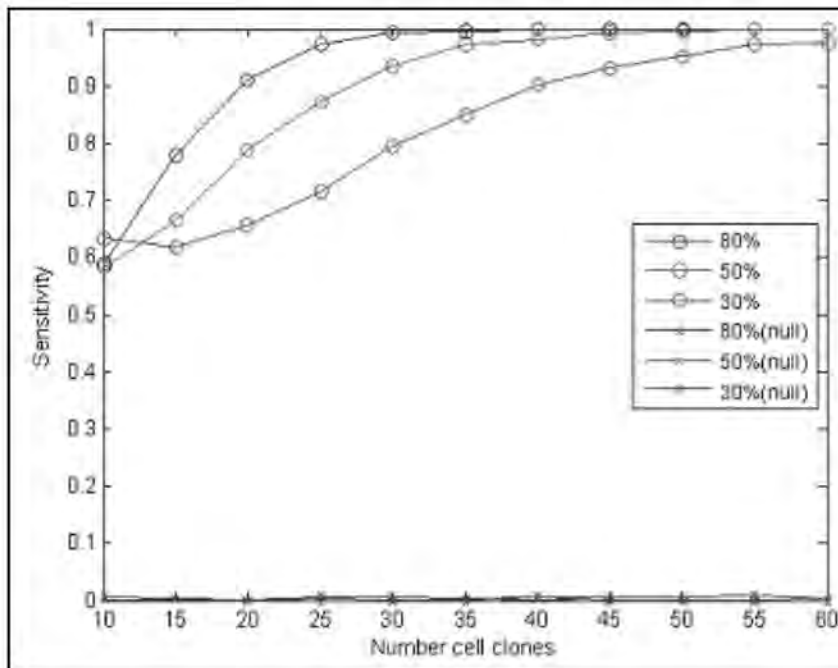


Figure 5-2: Sensitivity of detection of causality using assays of clonal altered cell lines, based on simulations involving 5 loci in a cis gene regulatory region. For locus **A/a**, allele **A** was assumed to cause its cis gene allele to be expressed at 2x the level of the **a** allele; other loci had no effect. Combinatorial cell populations were simulated in which all haplotypes occur with equal frequency whenever homologous recombination (HR) took place: HR efficiencies of 80%, 50%, and 30% were considered: Individual cells are isolated (number, X-axis), grown clonally, and assayed for genotype and gene expression level. Expression levels of the cis gene alleles phased with **A** and **a** were simulated subject to 20% Gaussian error. The **A/a** locus was identified as causal if gene expression level correlates with number of **A** alleles with $p < .001$ (corrected for 5 loci). "Sensitivity" = fraction of 1000 random simulations in which **A/a** was detected as causal. "null" lines: fraction of simulations in which non-causal locus **B/b** was identified as causal.

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

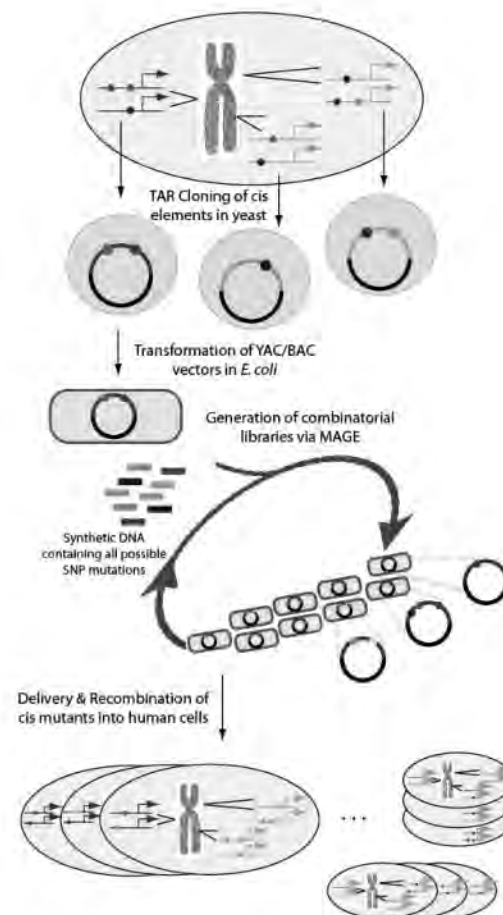


Figure 5.1.1-1: Illustration of MAGE-BAC/ZFN strategy for generating combinatorially modified human cells

Program Director/Principal Investigator (Last, First, Middle: Church, George M.

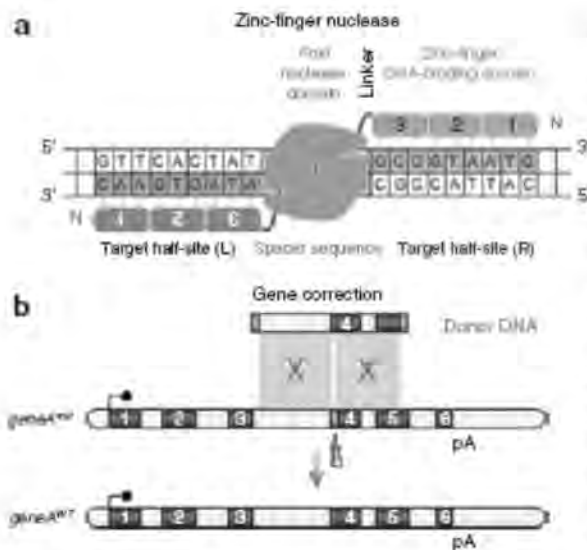
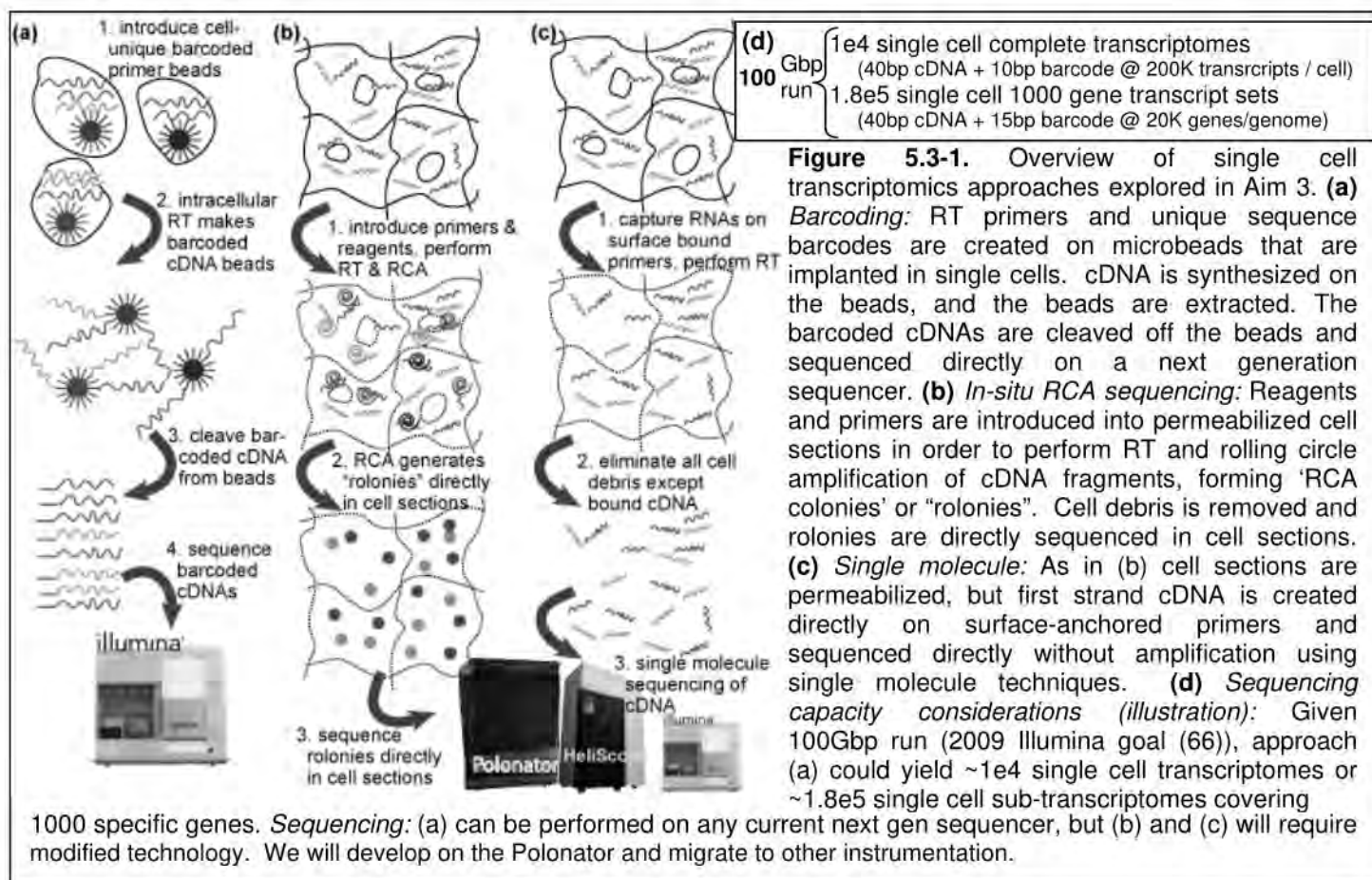


Figure 5.1.1-2. Engineered zinc-finger nucleases (ZFNs). (a) Architecture and application of ZFNs. A ZFN designed to create a DNA double-strand break (DSB) in the target locus comprises two monomer subunits. Each subunit contains three zinc-fingers (1-2-3), which recognize 9 base pairs within the full target site, and the *FokI* endonuclease domain (green). Dimerization activates the nuclease, cutting the DNA in the spacer sequence separating the target halfsites (L) and (R). ZFN subunits comprising four zinc-fingers that recognize 12 base pairs have also been developed. (b) ZFN-mediated gene disruption and correction by homologous recombination (HR). A DSB (yellow flash) is introduced by the ZFN into mutant allele A^{mut} of a gene. The presence of donor wild-type DNA drives DSB repair through HR vs. error-prone non-homologous end joining, yielding a functional wild-type allele geneA^{WT}. Rather than repair genes, CTCHGV will use ZFNs to mutate cis gene regulatory regions in order to determine their causal role in allele-specific expression. Figure adapted from Cathomen, 2008 (see References).

Program Director/Principal Investigator (Last, First, Middle: Church, George M.



Program Director/Principal Investigator (Last, First, Middle): Church, George M.

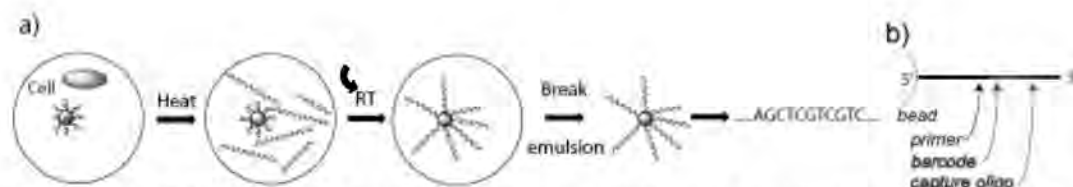


Figure 5.3.1.1-1. single cell mRNA capture and barcoding in emulsion. a) A water droplet containing a single cell and bead. The cell is lysed by heat, bound capture oligos bind to mRNA target sequences, reverse transcriptase is introduced and places the cDNA onto the bead, the emulsion is broken, and the cDNAs collected and sequenced. b) depiction of a bead-bound oligo. 5' primer region (black) allows subsequent amplification of captured cDNA, a bar-code (red) identifies transcripts belonging to the same cell, and the capture oligo (blue) captures the mRNA.

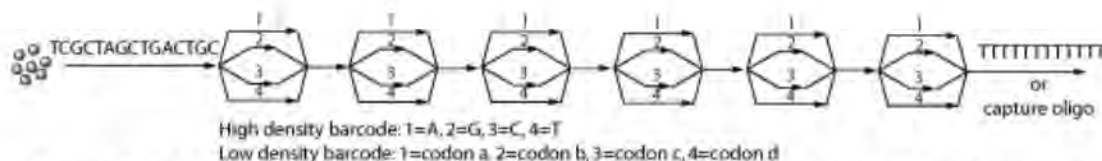


Figure 5.3.1.1-2. Split pooled synthesis of bar-coded oligonucleotides. 5' fixed sequence is grown on the beads. To apply the bar-code, beads are split into four pots and a reagent (single nucleotide or nucleotide triplet) is added to each. The beads are mixed and the process repeated. After addition of the bar-code, beads are pooled and poly-T or capture oligos are added.

Program Director/Principal Investigator (Last, First, Middle: Church, George M.

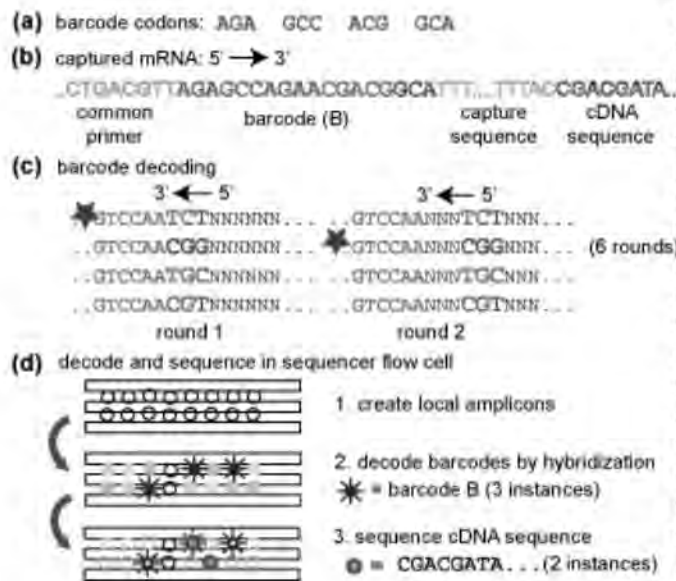


Figure 5.3.1.1-3. Illustration of low density barcodes. (a) Barcode codons. (b) Barcoded cDNA captured from single cell using undirected sequencing capture sequence polyT-AC. All cDNAs with this barcode ("B") come from the same cell. (c) Hybridization rounds used to decode barcodes. Starred primers correspond to barcode B. (d) Assignment of cDNA to cell using barcodes followed by sequencing of cDNAs. Local amplicons of the cDNAs are generated for sequencing. Barcode probing as in (c) assigns amplicons to cells. Sequencing from capture sequence or adaptor identifies cDNA. In this illustration, 3 features are assigned to barcode B, and two cDNAs have sequence of cDNA in (b), one of which is in cell B.

Program Director/Principal Investigator (Last, First, Middle: Church, George M.

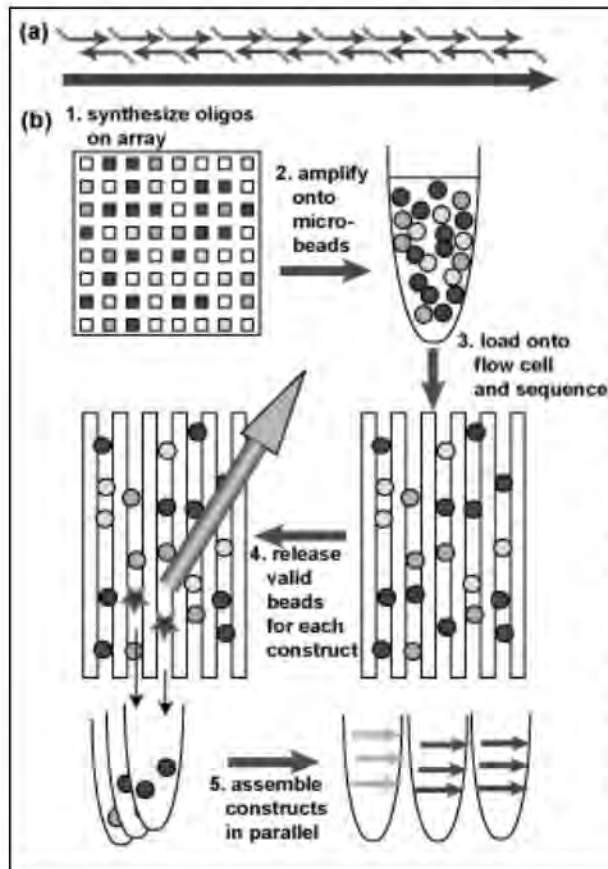


Figure 5.4.1-1. Schematic for one way of integrating DNA sequencing and synthesis for high-throughput reduced-error synthesis of large constructs. **(a)** Large DNA construct is analyzed into oligos with appropriate overlaps, uniqueness, Tms, as needed. **(b)** Processing pathway from synthesis of oligos on array for multiple constructs (represented by different colors) to multiplex synthesis. Amplification in (2) is illustrated as emulsion PCR as in (155). Microbeads are loaded onto flow cell using light-labile chemical attachments (see text) and sequenced on the flow cell (3). For each construct, light is directed to microbeads with sequence-validated oligos for the construct for release and capture (4). Assembly of all constructs then proceeds in parallel (5).

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

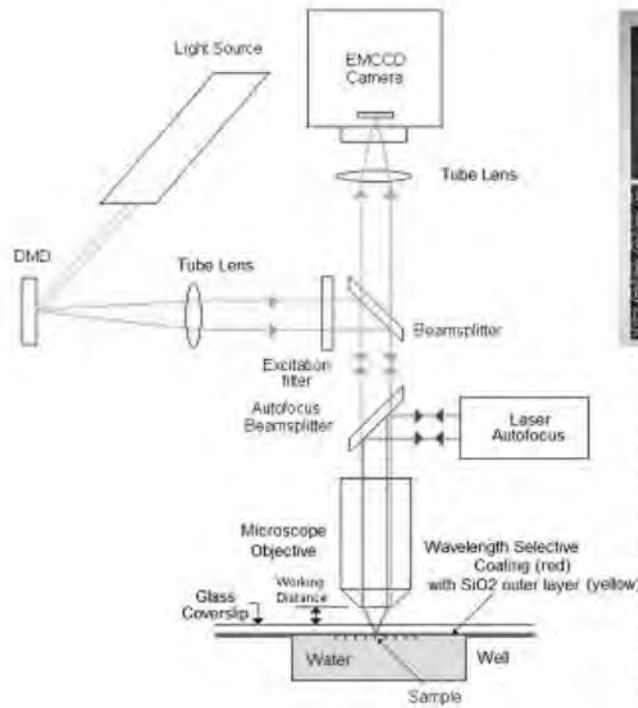


Figure 5.4.1-2: Integration of Digital Micro-mirror Device (DMD) array with the Polonator optical components. *Left:* Schematization of the optical path of light for DMD array control allowing selective release of cells from the Polonator flow cell. *Right:* Scanning Electron Microscope image of DMD mirrors and pivoting structure (Texas Instruments).

Program Director/Principal Investigator (Last, First, Middle: Church, George M.

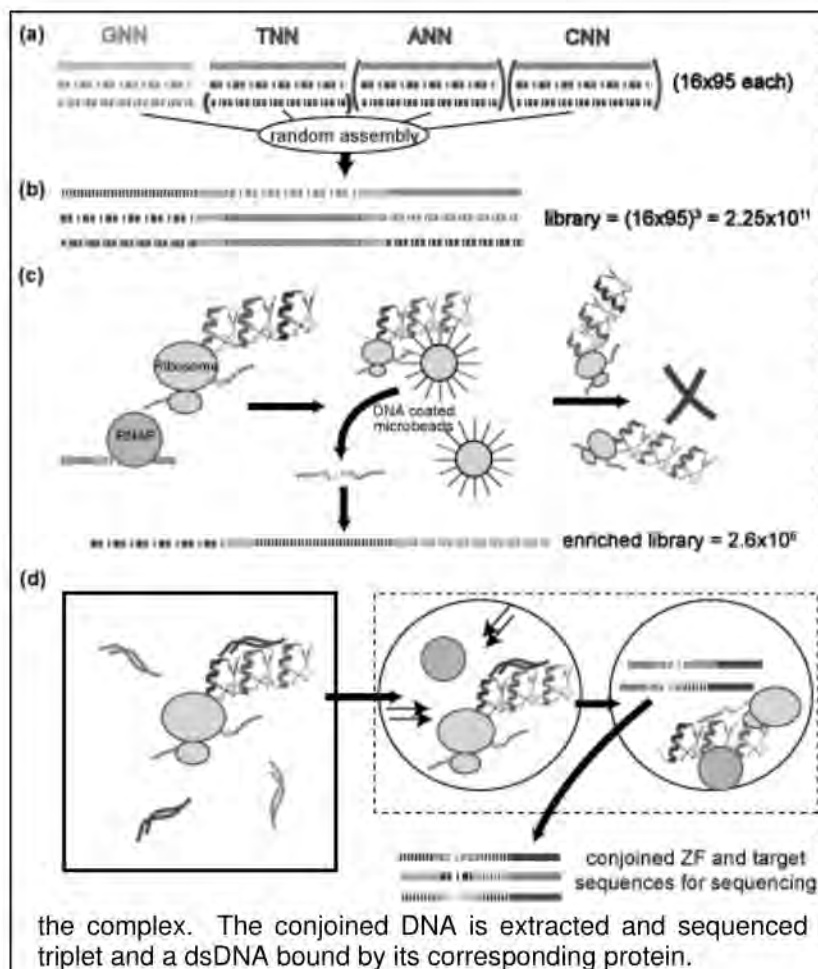
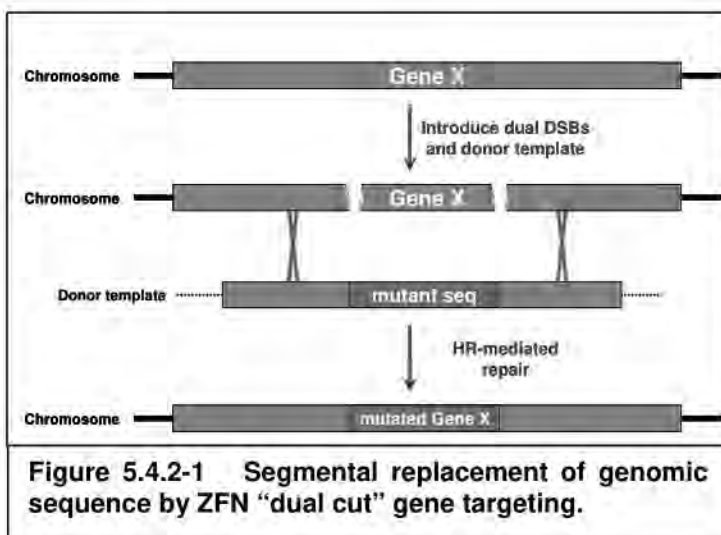


Figure 5.4.2-1. Strategy for ZFN improvement. (a) Completion of OPEN pools for all 3bp DNA sites. Each colored, dashed line represents a coding sequence specifying a zinc finger domain that recognizes a DNA triplet. Stacks of domains in parenthesis remain to be completed (section 5.4.2 i.A). (b) The pools of (a) will be combined into a library that codes all possible three zinc-finger OPEN pool combinations (section 5.4.2 i.B). (c) Ribosome display is used on the library of (b) to express all three zinc finger domains (left), and the RNA/Ribosome/ zinc finger triplet complexes are then exposed to microbeads coated with all double stranded DNA (dsDNA) 10mers (center). RNA extracted from beads will be used to generate the enriched library (bottom), while complexes that do not attach to beads will be washed away (right). (section 5.4.2 i.B) (d) Ribosome display is performed on the enriched library from (c) (box on right), and the RNA/Ribosome/zinc finger triplets exposed again to all dsDNA 10mers. RNA/Ribosome/zinc finger triplet/dsDNA complexes are purified, and an emulsion is created that contains 1 complex per compartment along with primers and enzymes (left circle of dashed box) for RT and PCR. RT, and then short overlap extension PCR (see section 5.4.2 i.C) are conducted in the emulsion (right circle of dashed box) to conjoin the bound dsDNA with cDNA from (bottom). Each sequence specifies a zinc finger

Program Director/Principal Investigator (Last, First, Middle): Church, George M.



Program Director/Principal Investigator (Last, First, Middle: Church, George M.

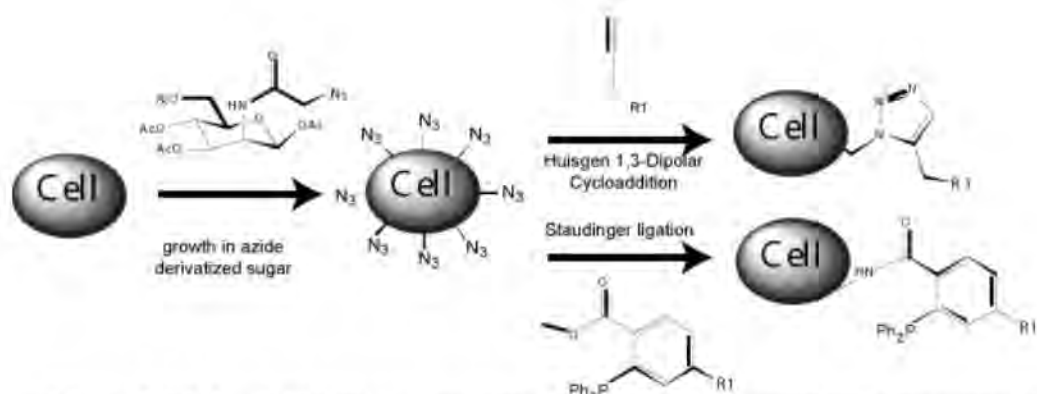


Figure 5.4.3-1. "click" chemistry capture of cells. Cells are grown in the presence of the azido sugar, which is displayed on the surface of cells. Azide groups undergo Huisgen cycloadditions or Staudinger ligations (148) to hetero-bifunctional linkers or solid surfaces. Table 5.4.3-1 describes combinations of hetero-bifunctional linkers with various functional groups (R_1), surface coatings, and crosslinking agents.

SUMMARY STATEMENT
(Privileged Communication)

PROGRAM CONTACT:
JEFFERY SCHLOSS PH.D.
301-496-7531
schlossj@mail.nih.gov

Release Date: 12/07/2009
Revised Date: 09/03/2010

Application Number: 1 P50 HG005550-01

Principal Investigator

CHURCH, GEORGE M PHD

Applicant Organization: HARVARD UNIVERSITY (MEDICAL SCHOOL)

Review Group: GNOM-G (J1)
Genome Research Review Committee

Meeting Date: 11/05/2009
Council: JAN 2010
Requested Start: 04/01/2010

RFA/PA: PAR08-094
PCC: X7JS

Dual IC(s): MH

Project Title: Causal Transcriptional Consequences of Human Genetic Variation

SRG Action: Impact/Priority Score:

Human Subjects:

Animal Subjects:

Children:

Project Year	Direct Costs Requested
1	2,800,000
2	2,308,280
3	2,316,808
4	2,325,592
5	2,334,640
TOTAL	12,085,320

**Estimated
Total Cost**

ADMINISTRATIVE BUDGET NOTE: The budget shown is the requested budget and has not been adjusted to reflect any recommendations made by reviewers. If an award is planned, the costs will be calculated by Institute grants management staff based on the recommendations outlined below in the COMMITTEE BUDGET RECOMMENDATIONS section.

1 P50 HG005550-01
CHURCH, G

2

GNOM-G (J1)

REVISED

1P50HG005550-01 CHURCH, GEORGE

RESUME AND SUMMARY OF DISCUSSION: This application was submitted in response to an NHGRI Program Announcement for Centers of Excellence in Genomic Science (CEGS). This P50 program is designed to support multi-investigator, interdisciplinary centers in the development of highly innovative genomic approaches, concepts, and technologies that substantially advance the state of the art in the study of a biological problem. In addition, the Center must have a training component that leverages the strengths of the CEGS.

This P50 application for the Center for Transcriptional Consequences of Human Genetic Variation proposes development of methods for studying the role of genetic variation in the control of gene transcription, particularly for those variants in non-coding regions. Engineered gene regulatory regions will be used to target 1000 genes to determine which variations causally control gene transcription. Human induced pluripotent stem cells (iPS) will be analyzed to determine the variations that affect specific cell types. Methods for transcriptome analysis in millions of single cells simultaneously will be developed enabling analysis in complex human tissue. The review of the proposal included an applicant interview with Drs. George Church, John Aach, and Keith Joung. Additional key personnel participating by telephone included Drs. Jehyuk Lee, Yveta Masarova, Sasha Wait Zaranek, and Kun Zhang.

reviewers' comments

1 P50 HG005550-01
CHURCH, G

3

GNOM-G (J1)

reviewers' comments

The comments in the CRITIQUE section were prepared by the reviewers assigned to this application and are provided without significant modification or editing by staff. The RESUME AND SUMMARY OF DISCUSSION section documents the final outcome of the evaluation by reviewers and is the basis for the assigned priority score.

DESCRIPTION (provided by applicant): The Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV) will develop innovative and powerful genetic engineering methods and use them to identify genetic variations that causally control gene transcription levels. Genome Wide Association Studies (GWAS) find many variations associated with disease and other phenotypes, but the variations that may actually cause these conditions are hard to identify because nearby variations in the same haplotype blocks consistently co-occur with them in human populations, so that specifically causative ones cannot be distinguished. About 95% of GWAS variations are not in gene coding regions, and many of these presumably associate with altered gene expression levels. CTCHGV will identify the variations that directly control gene expression by engineering precise combinations of changes to gene regulatory regions that break down the haplotype blocks, allowing each variations' effect on gene expression to be discerned independently of the others. To perform this analysis, CTCHGV will extract ~100kbp gene regulatory regions from human cell samples, create precise variations in them in *E. coli*, and re-introduce the altered regions back into human cells, using zinc finger nucleases (ZFNs) to efficiently induce recombination. CTCHGV will target 1000 genes for this analysis (Aim 1), and will use human induced Pluripotent Stem cells (iPS) to study the effects of variations in diverse human cell types (Aim 2). To explore the effects of variations in complex human tissues, CTCHGV will develop methods of measuring gene expression at transcriptome-wide levels in many single cells, including in situ in structured tissues (Aim 3). Finally, CTCHGV will develop novel advanced technologies that integrate DNA sequencing and synthesis to construct thousands of large DNA constructs from oligonucleotides, that enable very precise targeting and highly efficient performance of ZFNs, and that enable cells to be sorted on the basis of morphology as well as fluorescence and labeling (Aim 4). CTCHGV will also develop direct oligo-mediated engineering of human cells, and create "marked allele" iPS that will enable easy ascertainment of complete exon distributions for many pairs of gene alleles in many cell types.

RELEVANCE: CTCHGV methods will yield precise knowledge of effects of human genetic variations on gene expression that will both refine and go beyond GWAS-derived associations between non-coding variations and disease. Powerful new CTCHGV genetic engineering methods will directly enable gene therapy. CTCHGV iPS and single-cell transcriptome technologies will increase understanding of diverse and complex human tissues.

CRITIQUE 1:

reviewers' comments

VIZ00099784

1 P50 HG005550-01
CHURCH, G

4

GNOM-G (J1)

reviewers' comments

**Overall Impact:
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

**1. Significance:
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

**2. Investigator(s):
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

**3. Innovation:
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

**4. Approach:
Strengths**

reviewers' comments

VIZ00099785

1 P50 HG005550-01
CHURCH, G

5

GNOM-G (J1)

reviewers' comments

Weaknesses

reviewers' comments

**5. Environment:
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

Budget and Period of Support:

reviewers' comments

Resource Sharing Plans:

reviewers' comments

CRITIQUE 2:

reviewers' comments

**Overall Impact:
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

**1. Significance:
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

1 P50 HG005550-01
CHURCH, G

6

GNOM-G (J1)

**2. Investigator(s):
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

**3. Innovation:
Strengths**

reviewers' comments

Weaknesses

**4. Approach:
Strengths**

reviewers' comments

Weaknesses

reviewers' comments

**5. Environment:
Strengths**

reviewers' comments

Weaknesses

Protections for Human Subjects:

reviewers' comments

Inclusion of Women, Minorities and Children:

reviewers' comments

Vertebrate Animals:

reviewers' comments

Biohazards:

reviewers' comments

Budget and Period of Support:

reviewers' comments

Select Agents:

reviewers' comments

Resource Sharing Plans:

reviewers' comments

VIZ00099787

1 P50 HG005550-01
CHURCH, G

7

GNOM-G (J1)

CRITIQUE 3:

Overall Impact: Strengths

reviewers' comments

Weaknesses

reviewers' comments

1. Significance: Strengths

reviewers' comments

Weaknesses

reviewers' comments

2. Investigator(s): Strengths

reviewers' comments

Weaknesses

reviewers' comments

3. Innovation: Strengths

reviewers' comments

VIZ00099788

1 P50 HG005550-01
CHURCH, G

8

GNOM-G (J1)

reviewers' comments

Weaknesses

reviewers' comments

4. Approach: Strengths

reviewers' comments

Weaknesses

reviewers' comments

5. Environment: Strengths

reviewers' comments

Weaknesses

reviewers' comments

Biohazards:

reviewers' comments

Budget and Period of Support:

reviewers' comments

Resource Sharing Plans:

reviewers' comments

CRITIQUE 4:

Environment: Strengths

reviewers' comments

1 P50 HG005550-01
CHURCH, G

9

GNOM-G (J1)

Weaknesses

reviewers' comments

1. Significance

Strengths

reviewers' comments

Weaknesses

reviewers' comments

2. Investigator(s)

Strengths

reviewers' comments

Weaknesses

reviewers' comments

3. Innovation

Strengths

reviewers' comments

Weaknesses

reviewers' comments

4. Approach

Strengths

reviewers' comments

Weaknesses

reviewers' comments

5. Environment

Strengths

reviewers' comments

Weaknesses

reviewers' comments

Protections for Human Subjects

reviewers' comments

VIZ00099790

1 P50 HG005550-01
CHURCH, G

10

GNOM-G (J1)

Inclusion of Women, Minorities and Children Applicable Only for Human Subjects Research

reviewers' comments

Vertebrate Animals

reviewers' comments

Budget and Period of Support

reviewers' comments

Resource Sharing Plans

reviewers' comments

THE FOLLOWING RESUME SECTIONS WERE PREPARED BY THE SCIENTIFIC REVIEW OFFICER TO SUMMARIZE THE OUTCOME OF DISCUSSIONS OF THE REVIEW COMMITTEE ON THE FOLLOWING ISSUES:

PROTECTION OF HUMAN SUBJECTS (Resume): reviewers' comments

reviewers' comments

INCLUSION OF WOMEN PLAN (Resume): reviewers' comments

INCLUSION OF MINORITIES PLAN (Resume): reviewers' comments

INCLUSION OF CHILDREN PLAN (Resume): reviewers' comments

TRAINING: reviewers' comments

reviewers' comments

MANAGEMENT PLAN: reviewers' comments

reviewers' comments

DATA AND RESOURCE SHARING: reviewers' comments

reviewers' comments

COMMITTEE BUDGET RECOMMENDATIONS: reviewers' comments

reviewers' comments

SCIENTIFIC REVIEW OFFICER'S NOTES: reviewers' comments

reviewers' comments

REVISED SUMMARY STATEMENT: reviewers' comments

reviewers' comments

REVISED SUMMARY STATEMENT: (09/03/2010) reviewers' comments

reviewers' comments

VIZ00099791

1 P50 HG005550-01
CHURCH, G

11

GNOM-G (J1)

EVALUATION OF THE CHURCH CEGS MINORITY ACTION PLAN
(March 24, 2010)

SRO Administrative Note: The Minority Action Plan (MAP) associated with this application was reviewed on March 23rd 2010. The review was conducted by a Special Emphasis Panel (SEP) and is included in this summary statement in order to keep it as part of the official record in the NIH database system. The evaluations and recommendations of the SEP were limited to the activities of the MAP and did not contribute to the priority score that was voted to the overall application. However, in the event that an award is contemplated for this application, NHGRI will not fund the parent grant until an approved MAP has been developed by the applicant. The Direct Cost Requested shown below, are included in the overall budget figures found on the face page of this summary statement.

<u>Project Year</u>	<u>Direct Costs Requested</u>
Year 1	295,880
Year 2	247,169
Year 3	253,740
Year 4	260,048
Year 5	266,544

MINORITY ACTION PLAN ASSESSMENT:

reviewers' comments

The PI is Dr. George Church and the MAP component is part of the recently reviewed Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV) Center of Excellence in Genomic Sciences (CEGS) at Harvard. This MAP application builds on the MAP component of Dr. Church's previous Molecular and Genomic Imaging (MGI) CEGS program and responds to the previous MAP review recommendations.

The MAP proposes to continue the work of the MGI CEGS MAP (2005-2009) with the aim to support 25 undergraduate Underrepresented Minorities (URMs) (5/year) through summer research experiences; the undergraduate students will work alongside of CTCHGV graduate students and post doctoral fellows. In addition, the MAP proposes to support 2 URM post doctoral fellows. Two focused activities are proposed. Activity 1 is the CTCHGV summer undergraduate research experiences (SURE) component and activity 2 is the CTCHGV MAP Post doctoral component. To carry out the two activities the application has five specific aims. Aim 1 will recruit 5 URM undergraduates per year to participate in mentored summer research experiences in a CTCHGV lab. Aim 2 will recruit one new URM post-doc per year for two year periods to pursue a mentored project in a CTCHGV lab. Aim 3 will provide mentoring to enable both undergraduate and post-doc URMs to move to the next steps of their science careers. Aim 4 will improve and formalize MAP milestones and processes for recruiting, mentoring, data gathering and analyzing, and tracking and evaluating trainees and program performance. Initial milestone targets will be 50% of CTCHGV MAP summer undergraduate interns will apply to graduate school programs in the biological or biomedical sciences and will proceed to an advanced degree (Ph.D., M.D., or M.D. /Ph.D.) and 95% of CTCHGV MAP post-docs will transition to a career in the biological and biomedical sciences in academia, industry, or government. Aim 5 will improve coordination with other MAP programs and services, with focus on partnering with similar programs in the Boston area.

reviewers' comments

VIZ00099792

1 P50 HG005550-01
CHURCH, G

12

GNOM-G (J1)

reviewers' comments

KMcK

ASSESSOR 1

1. Assessment of Specific Activities

Activity 1: Summer Undergraduate Research Experience

reviewers' comments

VIZ00099793

1 P50 HG005550-01
CHURCH, G

13

GNOM-G (J1)

reviewers' comments

Activity 2: Post doc Program

reviewers' comments

VIZ00099794

1 P50 HG005550-01
CHURCH, G

14

GNOM-G (J1)

reviewers' comments

2. Evaluation of Past Performance

reviewers' comments

3. Assessment of Overall Action Plan

reviewers' comments

ASSESSOR 2

1. Assessment of Specific Activities

Activity 1: CTCHGV CEGS Proposed Minority Action Plan-Undergraduate Program

- What are the strengths and weaknesses of the activity?

reviewers' comments

1 P50 HG005550-01
CHURCH, G

15

GNOM-G (J1)

- How could the activity be improved? What elements should be included in this activity to make it an effective program?

reviewers' comments

- How will the activity facilitate participants moving to the next phase of their educational or career program?

reviewers' comments

- Are the milestones appropriate? If not, how can they be refined?

reviewers' comments

- Is the evaluation component appropriate? If not, how can they be refined?

reviewers' comments

- Are individuals with the right expertise involved in the development and management of the program?

reviewers' comments

- Will this activity facilitate the long-term goal of the NHGRI Minority Action Plan?

reviewers' comments

- Is the budget appropriate?

reviewers' comments

ACTIVITY RATING (check one):

reviewers' comments

Assessment of Specific Activities

Activity 2: CTCHGV CEGS Proposed Minority Action Plan- Post-Doc Program

- What are the strengths and weaknesses of the activity?

reviewers' comments

- How could the activity be improved? What elements should be included in this activity to make it an effective program?

reviewers' comments

1 P50 HG005550-01
CHURCH, G

16

GNOM-G (J1)

reviewers' comments

- How will the activity facilitate participants moving to the next phase of their educational or career program?

reviewers' comments

- Are the milestones appropriate? If not, how can they be refined?

reviewers' comments

- Is the evaluation component appropriate? If not, how can they be refined?

reviewers' comments

- Are individuals with the right expertise involved in the development and management of the program?

reviewers' comments

- Will this activity facilitate the long-term goal of the NHGRI Minority Action Plan?

reviewers' comments

- Is the budget appropriate?

reviewers' comments

ACTIVITY RATING (check one):

reviewers' comments

2. For Competitive Renewals, evaluate past performance

- Did the current MAP provide an adequate evaluation of the program's performance during the previous funding period? Was there adequate progress?

reviewers' comments

- Were there any program changes made as a result of the evaluation outcome and were they appropriate?

reviewers' comments

- What evidence was presented showing that past participants were productive? Have any transitioned successfully to the next career level?

reviewers' comments

3. Assessment of Overall Action Plan

- Are the goals of the program clear and will the objectives and activities accomplish the goals?

reviewers' comments

VIZ00099797

1 P50 HG005550-01
CHURCH, G

17

GNOM-G (J1)

- How well is genomics integrated into planned activities?

reviewers' comments

- Does the PI have sufficient involvement in the MAP program?

reviewers' comments

reviewers' comments

- If the MAP includes participation in an ongoing activity designed and managed by others, is there adequate involvement and participation of the PI and MAP personnel in the activity(s) to provide "added-value?"

reviewers' comments

- Is the level of funding for the proposed training activities commensurate with the requested level of funding for the entire project?

reviewers' comments

- Summarize your evaluation of overall Minority Action Plan.

reviewers' comments

reviewers' comments

OVERALL RATING (check one):

reviewers' comments

1 P50 HG005550-01
CHURCH, G

18

GNOM-G (J1)

ASSESSOR 3

Activity 1: CTCHGV Summer Undergraduate Research Experiences

STRENGTHS

reviewers' comments

CONCERNS/RECOMMENDATIONS

reviewers' comments

1 P50 HG005550-01
CHURCH, G

19

GNOM-G (J1)

reviewers' comments

Overall Activity Summary

reviewers' comments

Activity 2: Implementing a Post-doc Training Program

Strengths

reviewers' comments

VIZ00099800

1 P50 HG005550-01
CHURCH, G

20

GNOM-G (J1)

CONCERNS/RECOMMENDATIONS

reviewers' comments

Overall

reviewers' comments

OTHER POSITIVES

reviewers' comments

OTHER CONCERNS/RECOMMENDATIONS

reviewers' comments

ASSESSOR 4

1. Assessment of Specific Activities

VIZ00099801

1 P50 HG005550-01
CHURCH, G

21

GNOM-G (J1)

Activity 1: (Summer Undergraduate Research Experience (CTCHGV-SURE))

reviewers' comments

- How could the activity be improved? What elements should be included in this activity to make it an effective program?

reviewers' comments

- How will the activity facilitate participants moving to the next phase of their educational or career program?

reviewers' comments

- Are the milestones appropriate? If not, how can they be refined?

reviewers' comments

INSTITUTION

Principal Investigator

- Is the evaluation component appropriate? If not, how can they be refined?

reviewers' comments

- Are individuals with the right expertise involved in the development and management of the program?

1 P50 HG005550-01
CHURCH, G

22

GNOM-G (J1)

reviewers' comments

- Will this activity facilitate the long-term goal of the NHGRI Minority Action Plan?

reviewers' comments

- Is the budget appropriate?

reviewers' comments

ACTIVITY RATING (check one):

reviewers' comments

reviewers' comments

2. Assessment of Specific Activities

Activity 2: (CTCHGV-MAP Post Doc Program)

- What are the strengths and weaknesses of the activity?

reviewers' comments

- How could the activity be improved? What elements should be included in this activity to make it an effective program?

reviewers' comments

1 P50 HG005550-01
CHURCH, G

23

GNOM-G (J1)

- How will the activity facilitate participants moving to the next phase of their educational or career program?

reviewers' comments

- Are the milestones appropriate? If not, how can they be refined?

reviewers' comments

- Is the evaluation component appropriate? If not, how can they be refined?

reviewers' comments

- Are individuals with the right expertise involved in the development and management of the program?

reviewers' comments

INSTITUTION
Principal Investigator

reviewers' comments

- Will this activity facilitate the long-term goal of the NHGRI Minority Action Plan?

reviewers' comments

- Is the budget appropriate?

reviewers' comments

ACTIVITY RATING (check one):

reviewers' comments

3. For Competitive Renewals, evaluate past performance

- Did the current MAP provide an adequate evaluation of the program's performance during the

VIZ00099804

1 P50 HG005550-01
CHURCH, G

24

GNOM-G (J1)

previous funding period? Was there adequate progress?

reviewers' comments

• Were there any programs changes made as a result of the evaluation outcome and were they appropriate?

reviewers' comments

• What evidence was presented showing that past participants were productive? Have any transitioned successfully to the next career level?

reviewers' comments

4. Assessment of Overall Action Plan

• Are the goals of the program clear and will the objectives and activities accomplish the goals?

reviewers' comments

• How well is genomics integrated into planned activities?

reviewers' comments

• Does the PI have sufficient involvement in the MAP program?

reviewers' comments

• If the MAP includes participation in an ongoing activity designed and managed by others, is there adequate involvement and participation of the PI and MAP personnel in the activity(s) to provide "added-value?"

reviewers' comments

• How does the plan take advantage of the research infrastructure?

reviewers' comments

• Is the level of funding for the proposed training activities commensurate with the requested level of funding for the entire project?

reviewers' comments

• Summarize your evaluation of overall Minority Action Plan.

reviewers' comments

OVERALL RATING (check one):

reviewers' comments

MEETING ROSTER
National Human Genome Research Institute Special Emphasis Panel
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
MAP Review Panel
March 23, 2010

CHAIRPERSON

Jordan, Tuajuanda C PhD
Senior Program Officer
Division of Science Education Alliance

VIZ00099805

1 P50 HG005550-01
CHURCH, G

25

GNOM-G (J1)

Howard Hughes Medical Institute
Chevy Chase, MD 20815-6789

MEMBERS

Cantrarella, Marcia Young PhD
Associate Dean and Professor
Office of the Dean
School of Arts and Science
Hunter College
New York, NY 10021

Johnson, Justine
Assistant Director
Meyerhoff Graduate Fellows Program
Howard Hughes Medical Institute
University of Maryland, Baltimore County
Baltimore MD 21250

Whittington, Dawayne PhD
Consultant
Strategic Evaluations, Inc
Durham, NC 27707

SCIENTIFIC REVIEW ADMINISTRATOR

McKenney, Keith H. PhD
Scientific Review Administrator
Scientific Review Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, Md 20892

Consultants are required to absent themselves from the room during the review of any application if their presence would constitute or appear to constitute a conflict of interest.

NIH has modified its policy regarding the receipt of resubmissions (amended applications). See Guide Notice NOT-OD-10-080 at <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-10-080.html>.

The impact/priority score is calculated after discussion of an application by averaging the overall scores (1-9) given by all voting reviewers on the committee and multiplying by 10. The criterion scores are submitted prior to the meeting by the individual reviewers assigned to an application, and are not discussed specifically at the review meeting or calculated into the overall impact score. For details on the review process, see http://grants.nih.gov/grants/peer_review_process.htm#scoring.

MEETING ROSTER

**Genome Research Review Committee
National Human Genome Research Institute Initial Review Group
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
GNOM-G (J1) 1
November 05, 2009 - November 06, 2009**

CHAIRPERSON

NICKERSON, DEBORAH A, PHD
PROFESSOR
DEPARTMENT OF GENOME SCIENCES
UNIVERSITY OF WASHINGTON
SCHOOL OF MEDICINE
SEATTLE, WA 98195

MEMBERS

ARNOSTI, DAVID N., PHD
PROFESSOR
DEPARTMENT OF BIOCHEMISTRY
AND MOLECULAR BIOLOGY
MICHIGAN STATE UNIVERSITY
EAST LANSING, MI 488241319

BOEKE, JEF D, PHD
PROFESSOR
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS
THE JOHN HOPKINS UNIVERSITY
SCHOOL OF MEDICINE
BALTIMORE, MD 21205

FRASER, SCOTT E, PHD
PROFESSOR
DIVISION OF BIOLOGY
BECKMAN INSTITUTE
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CA 91125

FRAZER, KELLY A., PHD
PROFESSOR
DEPARTMENT OF PEDIATRICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
SCHOOL OF MEDICINE
LA JOLLA, CA 92093

KITTLES, RICK , PHD
ASSOCIATE PROFESSOR
SECTION OF GENETIC MEDICINE
DEPARTMENT OF MEDICINE
THE UNIVERSITY OF CHICAGO
CHICAGO, IL 60637

KWOK, PUI-YAN , MD, PHD
PROFESSOR
DEPARTMENT OF DERMATOLOGY
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
SAN FRANCISCO, CA 94143

NUSBAUM, HARRIS CHAD, PHD
CO-DIRECTOR
GENOME SEQUENCING AND ANALYSIS PROGRAMS
BROAD INSTITUTE
CAMBRIDGE, MA 02142

RZHETSKY, ANDREY , PHD
PROFESSOR
DEPARTMENT OF HUMAN GENETICS
INSTITUTE FOR GENOMICS AND SYSTEMS BIOLOGY
UNIVERSITY OF CHICAGO
CHICAGO, IL 60637

SJOLANDER, KIMMEN , PHD
ASSOCIATE PROFESSOR
DEPARTMENT OF BIOENGINEERING
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CA 947201762

WASSERMAN, WYETH W, PHD
PROFESSOR
DEPARTMENT OF MEDICAL GENETICS
CENTRE FOR MOLECULAR MEDICINE AND
THERAPEUTICS
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, BC V5Z 4H4
CANADA

WHITE, OWEN R., PHD
PROFESSOR
DEPARTMENT OF EPIDEMIOLOGY AND PREVENTIVE
MEDICINE
INSTITUTE FOR GENOME SCIENCES
UNIVERSITY OF MARYLAND SCHOOL OF MEDICINE
BALTIMORE, MD 21201

SCIENTIFIC REVIEW ADMINISTRATOR

NAKAMURA, KEN D., PHD
SCIENTIFIC REVIEW OFFICER
SCIENTIFIC REVIEW BRANCH
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
NATIONAL INSTITUTES OF HEALTH
BETHESDA, MD 20892

GRANTS TECHNICAL ASSISTANT

DEHAUT-COMBS, EDITH , MBA
GRANTS TECHNICAL ASSISTANT
SCIENTIFIC REVIEW BRANCH
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
BETHESDA, MD 20892

Consultants are required to absent themselves from the room during the review of any application if their presence would constitute or appear to constitute a conflict of interest.

Church CEGS

George Church. Co-PIs: George Daley, Keith Joung, Kun Zhang.

Harvard Medical School

READING AND WRITING GENOMES: CAUSAL CONSEQUENCES OF HUMAN GENETIC VARIATION

Aim 1: This project enables methods for highly multiplexed changes in genomes to test hypotheses flowing from genome sequencing and diverse human traits (especially cis-regulatory variations).

Aim 2: Automating establishment of human pluripotent stem cells (iPSC), diverse differentiated cell types and complex in vitro tissues. We are engineering heterozygosity in multiple exons of many genes to enable analysis of allele-specific transcription and splicing.

Aim 3: Exploring in situ RNA multiplex sequencing to monitor the impact of genetic variations in many RNAs and many cell types at once.

Aim 4: Developing methods for low-cost synthesis of long DNA constructs, efficient homologous recombination in human cells, and highly multiplexed single cell handling that enables sorting based on morphology.

Jeff Gole and Kun Zhang

Department of Bioengineering, University of California at San Diego

EXPERIMENTAL CONSTRUCTION OF FULLY PHASED DIPLOID GENOMES

Over 99% of genetic variations in human population are located outside protein coding sequences. Many phenotypic differences among human individuals are believed to be determined by such non-coding variants. However, identifying and characterizing causative variants in non-coding regions is challenging, since they often are located far away from the genes they regulate. We have developed an experimental method to construct fully phased diploid genomes, which will be used as a chassis to establish the end-to-end connectivity among all genetic variants across the entire chromosomes. The fully phased diploid genome will allow us to investigate the combinatorial effects of multiple cis-regulatory genetic variants on their target genes regardless of their linear distance on the chromosomes.

To construct fully phased diploid genome, we have adapted the polymerase cloning method to intact human chromosome molecules. We use polymerase cloning to derive 8-12 haploid genomic libraries for each individual, with each library representing a random subset of chromosomes. By random shotgun sequencing of these libraries, we can establish end-to-end haplotypes for every human chromosome. To develop a robust procedure for routine diploid genome sequencing, we have implemented the polymerase cloning method in microwell arrays, such that tens of thousands of reactions can be self-assembled and performed in parallel. The resulting amplicons can be screened by in situ fluorescent probing and retrieved with a microinjection system for shotgun sequencing. We expect to establish a robust and scalable haplotyping pipeline toward the end of this project.

Progress Report Scanning Cover Sheet

5P50HG005550-02

PI Name: **CHURCH, GEORGE**
Org: **HARVARD UNIVERSITY (MEDICAL SCHOOL)**
Start Date: **08/01/2011**
Snap: **N/A (NEEDS TO BE BOOKMARKED)**
Appl ID: **8141976**
Rec'd Date: **05/20/2011**

Form Approved Through 06/30/2012

OMB No. 0925-0001

Department of Health and Human Services
Public Health Services

Review Group	Type 5	Activity P50	Grant Number HG005550-02
--------------	-----------	-----------------	-----------------------------

Grant Progress Report

Total Project Period	
From: 09/13/2010	Through: 07/31/2015
Requested Budget Period	
From: 08/01/2011	Through: 07/31/2012

1. TITLE OF PROJECT

Causal Transcriptional Consequences of Human Genetic Variation

2a. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR

(Name and address, street, city, state, zip code)
George M. Church, PhD.
Harvard Medical School
Department of Genetics
77 Avenue Louis Pasteur, NRB 238
Boston, MA 02115

2b. E-MAIL ADDRESS

gmc@harvard.edu

2c. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT

Genetics

2d. MAJOR SUBDIVISION

Harvard Medical School

2e. Tel: 617-432-7562

Fax: 617-432-6513

3a. APPLICANT ORGANIZATION

(Name and address, street, city, state, zip code)
Harvard Medical School
25 Shattuck St., suite 509
Boston, MA 02115

3b. Tel: 617-432-1596

Fax: 617-432-2651

3c. DUNS: 047006379

MAY 20 2011

4. ENTITY IDENTIFICATION NUMBER
1042103580C56. HUMAN SUBJECTS ☐ No ☒ Yes6a. Research
Exempt☐ No ☒ YesIf Exempt ("Yes" in
6a).

Exemption No.

4

If Not Exempt ("No" in
6a).

IRB approval date

5. NAME, TITLE AND ADDRESS OF ADMINISTRATIVE OFFICIAL

Barbara Cevallos, Director, SPA
25 Shattuck St.
Boston, MA 02115

6b. Federal Wide Assurance No. FWA00007071

Tel: 617-432-1596

Fax: 617-432-2651

6c. NIH-Defined Phase III

Clinical Trial ☒ No ☐ Yes

E-MAIL: spa_award@hms.harvard.edu

7. VERTEBRATE ANIMALS ☒ No ☐ Yes

7a. If "Yes," IACUC approval Date

7b. Animal Welfare Assurance No. A3431-01

10. PROJECT/PERFORMANCE SITE(S)

Organizational Name: Harvard Medical School

DUNS: 047006379

8. COSTS REQUESTED FOR NEXT BUDGET PERIOD

8a. DIRECT \$2,559,065

8b. TOTAL \$3,809,825

Street 1: Harvard Medical School

Street 2:

9. INVENTIONS AND PATENTS ☐ No ☒ Yes

If "Yes," ☐ Previously Reported
☒ Not Previously Reported

City: Boston

County: Suffolk

State: MA

Province:

Country: United States

Zip/Postal Code: 02115

Congressional Districts: MA-008

11. NAME AND TITLE OF OFFICIAL SIGNING FOR APPLICANT ORGANIZATION (Item 13)

Deborah Good, Associate Director, Sponsored Programs Administration

TEL: 617-432-1596

FAX: 617-432-2651

E-MAIL: spa_award@hms.harvard.edu

12. Corrections to Page 1 Face Page

13. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE: I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.

SIGNATURE OF OFFICIAL NAMED IN

DATE

11. (In ink)

Signature

5/18/11

Contact Program Director/Principal Investigator: Church, George

2a. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR (Name and address, street, city, state, zip code) Kun Zhang, PhD. UCSD 9500 Gilman Dr MC 0412 La Jolla, CA 92093-0412	2b. E-MAIL ADDRESS kzhang@ucsd.edu 2c. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Bioengineering 2d. MAJOR SUBDIVISION General campus
--	--

2e. TELEPHONE AND FAX (Area code, number and extension)

TEL: 858-822-7876	FAX: 858-534-5722
-------------------	-------------------

2a. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR (Name and address, street, city, state, zip code) Keith J. Joung, PhD. Massachusetts General Hospital 13st Street Bldg 149 Charlestown, MA 02129	2b. E-MAIL ADDRESS jjoung@partners.org 2c. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Pathology 2d. MAJOR SUBDIVISION
---	---

2e. TELEPHONE AND FAX (Area code, number and extension)

TEL: 617-726-9462	FAX: 617-726-5684
-------------------	-------------------

2a. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR (Name and address, street, city, state, zip code) George Q. Daley, PhD. Children's Hospital Boston 300 Longwood Avenue Boston, MA 02115	2b. E-MAIL ADDRESS george.daley@childrens.harvard.edu 2c. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT Medicine 2d. MAJOR SUBDIVISION Hematology/Oncology
--	--

2e. TELEPHONE AND FAX (Area code, number and extension)

TEL: 617-919-2013	FAX: 617-730-0222
-------------------	-------------------

2a. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR (Name and address, street, city, state, zip code)	2b. E-MAIL ADDRESS 2c. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT 2d. MAJOR SUBDIVISION
--	---

2e. TELEPHONE AND FAX (Area code, number and extension)

TEL:	FAX:
------	------

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

DETAILED BUDGET FOR NEXT BUDGET PERIOD – DIRECT COSTS ONLY		FROM 8/1/11	THROUGH 7/31/12	GRANT NUMBER P50 HG005550-02			
List PERSONNEL (Applicant organization only) Use Cal, Acad, or Summer to Enter Months Devoted to Project Enter Dollar Amounts Requested (omit cents) for Salary Requested and Fringe Benefits							
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	SALARY REQUESTED	FRINGE BENEFITS	TOTALS
Church - MAIN	PD/PI				760,134	160,460	920,594
MAP					173,049	55,105	228,153
MGH: K Joung	see consortium						
Children's: G Daley	see consortium						
UCSD: Zhang	see consortium						
SUBTOTALS					933,183	215,564	1,148,747
CONSULTANT COSTS							
Main - consultants: 83,000							83,000
EQUIPMENT (Itemize)							
SUPPLIES (Itemize by category)							
MAIN: 367,454							
MAP: 25,000							
UCSD: see consortium							
MGH: see consortium							
CHB: see consortium							392,454
TRAVEL							
MAIN: 21,029 MAP: 15,000 SUBS: see consortium							36,029
INPATIENT CARE COSTS							
OUTPATIENT CARE COSTS							
ALTERATIONS AND RENOVATIONS (Itemize by category)							
OTHER EXPENSES (Itemize by category)							
MAIN: 318,192							
MAP: 11,810 SUBS: see consortium							330,002
SUBTOTAL DIRECT COSTS FOR NEXT BUDGET PERIOD							\$ 1,990,232
CONSORTIUM/CONTRACTUAL COSTS		DIRECT COSTS					339,869
CONSORTIUM/CONTRACTUAL COSTS		FACILITIES AND ADMINISTRATIVE COSTS					228,964
TOTAL DIRECT COSTS FOR NEXT BUDGET PERIOD (Item 8a, Face Page)							\$ 2,559,065

Program Director/Principal Investigator (Last, First, Middle):

Church, George M.

DETAILED BUDGET FOR NEXT BUDGET PERIOD - DIRECT COSTS ONLY					FROM 8/1/11	THROUGH 7/31/12	Grant Number P50 HG005550-02	
PERSONNEL (Applicant organization only)		Months Devoted to Project			INST. BASE SALARY	DOLLAR AMOUNT REQUESTED (omit cents)		
NAME	ROLE ON PROJECT	Cal. Mths	Acad. Mths	Summer Mths		SALARY REQUESTED	FRINGE BENEFITS	TOTAL
Church, George M.	PD/PI	EFFORT			Institutional Base Salary	71,892	20,417	92,309
John Aach	Senior Scientist					68,850	19,553	88,403
Yveta Masarova	Coordinator					38,031	16,924	54,954
Sara Vassallo	Technician					12,551	7,204	19,755
Richard Terry	Engineer					12,607	5,610	18,217
Francois Vigneault	PostDoc						No salary	requested
Jehyuk Lee	PostDoc					23,970	6,352	30,322
Michael Sismour	PostDoc					23,970	6,352	30,322
Adrian Briggs	PostDoc					5,000		5,000
Hamid Mukhtar	PostDoc					5,000		5,000
Sasha Wait Zaranek	PostDoc					10,656	2,824	13,480
Dan Mandel	PostDoc					39,400	10,441	49,841
Chris Gregg	PostDoc					39,400	10,441	49,841
Madeleine Ball	PostDoc					40,900	10,839	51,739
Jong Kim	PostDoc					45,960	12,179	58,139
Volker Busskamp	PostDoc					39,400	10,441	49,841
Prashant Mali	PostDoc					39,400	10,441	49,841
Yoav Mayshar	PostDoc					39,400	10,441	49,841
subtotal from pg 2	Grad students					203,748		203,748
						760,134	160,460	920,594
CONSULTANT COSTS: \$ 83,000								83,000
EQUIPMENT (Itemize):								
SUPPLIES (Itemize by category): Church lab: \$352,454 Computers: \$15,000								367,454
TRAVEL: Main: \$21,029								21,029
PATIENT CARE COSTS INPATIENT								
ALTERATIONS AND RENOVATIONS (Itemize by category)								
OTHER EXPENSES (Itemize by category)								
Tuition: \$32,536 (X Rios: \$12,048; K Robasky: \$8,488; Grad Program fee: \$12,000)								
Service contracts: \$23,794 Media & Glass washing: \$5,150								
Sequencing: \$220,000								
Publications: \$15,000								
Server charges (\$834.3/month): \$10,012								
Computer backup: \$11,700								318,192
SUBTOTAL DIRECT COSTS FOR NEXT BUDGET PERIOD								\$ 1,710,269
CONSORTIUM/CONTRACTUAL COSTS		DIRECT COSTS						
CONSORTIUM/CONTRACTUAL COSTS:		FACILITIES AND ADMIN COSTS:						
TOTAL DIRECT COSTS FOR NEXT BUDGET PERIOD (Item 8a, Face Page)								\$ 1,710,269

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

DETAILED BUDGET FOR INITIAL BUDGET PERIOD DIRECT COSTS ONLY						FROM 8/1/11	THROUGH 7/31/12	
PERSONNEL (<i>Applicant organization only</i>)		Months Devoted to Project			INST.BASE SALARY	DOLLAR AMOUNT REQUESTED (<i>omit cents</i>)		
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths		SALARY REQUESTED	FRINGE BENEFITS	TOTAL
Xavier Rios (Harvard)	Grad Stud	EFFORT			Institutional Base Salary	31,724		31,724
Kim Robasky (BU)	Grad Stud					33,187		33,187
Joyce Yang (Harvard)	Grad Stud					31,512		31,512
Luhan Yang (Harvard)	Grad Stud					31,512		31,512
Uri Laserson (MIT)	Grad Stud					45,213		45,213
Le Cong (Harvard)	Grad Stud					30,600		30,600
Daniel Goodman (MIT)	Grad Stud						No salary	requested
Subtotal – grad students					203,748		203,748	

MAP budget

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

DETAILED BUDGET FOR NEXT BUDGET PERIOD – DIRECT COSTS ONLY	FROM 8/1/11	THROUGH 7/31/12	GRANT NUMBER P50 HG005550-02
---	-----------------------	---------------------------	--

List PERSONNEL (Applicant organization only)

Use Cal, Acad, or Summer to Enter Months Devoted to Project

Enter Dollar Amounts Requested (omit cents) for Salary Requested and Fringe Benefits

Enter total funding requested (sum entry for salary requested and fringe benefits)							
NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	SALARY REQUESTED	FRINGE BENEFITS	TOTALS
Church, George	PD/PI						
Willie, Giraldd-Rosa	PostDoc	EFFORT			39,400	10,441	49,841
TBH	PostDoc				32,833	8,701	41,534
Lee Bitsoi	MAP prog Director				80,815	35,963	116,778
TBH	Summer student				4000		4000
TBH	Summer student				4000		4000
TBH	Summer student				4000		4000
TBH	Summer student				4000		4000
TBH	Summer student				4000		4000
SUBTOTALS					173,049	55,105	228,153

CONSULTANT COSTS

EQUIPMENT (Itemize)

SUPPLIES (Itemize by category)

Lab supplies : 23,000

Computers: 2,000

25,000

TRAVEL

Travel to conference: 6 trips (2,500)

15,000

INPATIENT CARE COSTS

OUTPATIENT CARE COSTS

ALTERATIONS AND RENOVATIONS (Itemize by category)

OTHER EXPENSES (Itemize by category)

Housing: (765/month): \$3,826

Recruiting- advertisement: \$500

Candidates – interview expenses: \$4,007

Lunch meetings: \$400

GRE course: \$2,477

Mentoring event: \$600

11,810

SUBTOTAL DIRECT COSTS FOR NEXT BUDGET PERIOD**\$ 279,963**

CONSORTIUM/CONTRACTUAL COSTS

DIRECT COSTS

CONSORTIUM/CONTRACTUAL COSTS

FACILITIES AND ADMINISTRATIVE COSTS

TOTAL DIRECT COSTS FOR NEXT BUDGET PERIOD (Item 8a, Face Page)**\$ 279,963**

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

BUDGET JUSTIFICATION

 GRANT NUMBER
 P50 HG005550

Provide a detailed budget justification for those line items and amounts that represent a significant change from that previously recommended. Use continuation pages if necessary.

Main budget - no significant changes

MAP budget - no significant changes

CURRENT BUDGET PERIOD

 FROM
 9/13/2010

 THROUGH
 7/31/2011

Explain any estimated unobligated balance (including prior year carryover) that is greater than 25% of the current year's total budget.

Main project under direction of Dr. Church - unobligated balance - expected at the end of the project period. The Center began operations immediately after the award was granted, however the delays in set up of an account and shorter budget period resulted in more than 25% unspent funds

MAP project - less than 25%

Consortium Agreement:

less than 25% unobligated balance

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

BIOGRAPHICAL SKETCH

NAME Church, George, M., PhD.		POSITION TITLE	
eRA COMMONS USER NAME (credential, e.g., agency login) eRA Commons User Name		Professor	
EDUCATION/TRAINING			
INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
Duke University, Durham, NC	B.A.	1974	Zoology & Chem.
Harvard University, Cambridge, MA	PhD.	1984	Biochem. & Mol. Biol.

A. Personal Statement:

The goal of our part of the proposed research is to develop enabling technology for human genome sequencing and interpretation. Specifically, we plan to focus on large-scale data sets and community software for integrating omic, environmental and trait data. From my thesis to the present my group and I have pioneered many of the 2nd & 3rd-generation sequencing including pore, EM, polymerase, and ligase technologies [1,10] and applications to miRNAs[5], CpG mutations[6], drug resistance[2], RNA-editing[4] and RNA Allelotyping[8]. Computation has played huge role in nearly all paper have this laboratory. Translation of technologies into clinical and commercial sector has been a priority (advising and/or licensing to 15 nextgen sequencing companies, BGmedicine, Pharmorx, and Knome). I have also developed ELSI proposals[7] which deal with the challenge of making data and cells broadly available without overpromising on privacy. This lead to PersonalGenomes.org which has over 15,000 volunteers and a growing international consortium.

B. Positions and Honors:

1984 Scientist, Biogen Research Corporation, Cambridge, MA
 1985-1986 Research Fellow, Anatomy, Univ. Calif., San Francisco, CA
 1986-1998 Assistant/Associate Professor of Genetics, Harvard Medical School, Boston, MA
 1997-present Director of the Lipper Center for Computational Genetics, Boston, MA
 1998-present Professor of Genetics, Harvard Medical School, Boston, MA
 2002-present Director of the Harvard/MIT DOE Genomes-to-Life Center
 2004-present Director of the Harvard/MIT/WashU NHGRI CEGS
 2006-present Senior Associate of Broad Inst. of Harvard & MIT (1990 Genome Center Co-founder)

Honors, Awards, & Scientific Memberships:

1974-1975 National Science Foundation Predoctoral Fellow
 1985-1986 Life Sciences Research Foundation Fellow
 1976 National Science Foundation Program Project Grant Review Committee
 1986-1997 Howard Hughes Medical Institute
 1988,1992,1994 Department of Energy Genome Project Grant Review Committee
 1990 NIH Genome Study Section Grant Review
 1990 Co-founder of MIT, Stanford, & GTC Genome Sequencing Centers
 1994-1997 National Center for Human Genome Research Review Committee
 2001-present NIH BISTI, Pioneer, grant review committees, NHLBI BEE, NAS committees
 Editorial Boards Nature/EMBO-MSB, Genome Biology, Omics, BioMedNet
 Scientific Boards: LS9, 23andme, Knome, Genomatica, JouleBio, CompleteGenomics, Sigma-Aldrich, Halcyon
 2008 World Economic Forum Technology Pioneer Awards (LS9 & 23andme)
 2009 American Society for Microbiology Biotechnology Research Award

C. Selected peer-reviewed publications. (also see <http://arep.med.harvard.edu>)

1. Drmanac R, et al. (2010) Human Genome Sequencing Using Unchained Base Reads on Self-assembling DNA Nanoarrays. Science 2010 Jan 1;327(5961):78-81. Epub 2009 Nov 5; Supplement. PMID: 19892942
 2. Sommer MO, Dantas G, Church GM. (2009) Functional Characterization of the Antibiotic Resistance Reservoir in the Human Microflora. Science Aug 28; 325 (5944) 1128-31. PMID: 19713526

Program Director / Principal Investigator: Church, George M.

3. Kim JI, et al. (2009) A highly annotated whole genome sequence of a Korean Individual. Nature Jul 8; PMID: 19587683.
4. Li JB, Levanon EY, Yoon J-K, Aach J, Xie B, LeProust E, Zhang K, Gao Y, Church GM (2009) Genome-wide Identification of Human RNA Editing Sites by Massively Parallel DNA Capturing and Sequencing. Science. Jun 14; 324(5931):1210-3. PMID: 19478186
5. Vigneault F, Sismour AM, Church GM. (Sep 2008) Efficient microRNA capture and barcoding via enzymatic oligonucleotide adenylation. Nature Methods 5, 777 - 779. PMID: 18711363
6. Li JB, Gao Y, Aach J, Zhang K, Kryukov GV, Xie B, Ahlford A, Yoon J-K, Rosenbaum AM, Zaranek AW, LeProust E, Sunyaev SR, Church GM (2009) Multiplex padlock capturing and sequencing reveal human hypermutable CpG variations. Genome Research Sep;19(9):1606-15. PMID: 19525355.
7. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. From genetic privacy to open consent. Nat Rev Genet. 2008 May;9(5):406-11. PMID: 18379574
8. Zhang K, Li JB, Gao Y, Egli D, Xie B, Lee JH, Aach J, LeProust E, Eggan K, Church GM (2009) Digital RNA Allelotyping Reveals Tissue-specific and Allele-specific Gene Expression In Human. Nature Methods Aug;6(8):613-8. PMID: 19620972 PMC2742772
9. Zhang K, Zhu J, Shendure J, Porreca GJ, Aach JD, Mitra RD, Church GM (2006) Long-range polony haplotyping of individual human chromosome molecules. Nature Methods Aug;6(8):613-8. PMID: 19620972 PMC2742772
10. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome Science. 2005 Sep 9;309(5741):1728-32. PMID: 16081699

D. Research Support:**Ongoing:**

DE-FG02-02ER63445 (GTL) 2/01/03 – 11/30/11
 PI: George Church
 Title: Microbial Ecology, Proteogenomics & Computational Optima

P50 HG005550 (CEGS) 9/13/10-07/31/2015
 NIH- NHGRI
 PI: George Church
 Title: Center for Causal Transcriptional Consequences of Human Genetic Variation

SA5283-11210 (NSF) 7/01/10 – 6/30/14
 PI: Jay Keasling (UC Berkeley)
 Title: Synthetic Biology Engineering Research Center (SynBERC)

RO1 HL 094963- 01 (NHLBI) 9/30/08 - 6/30/11 (NCE)
 NIH - NHLBI
 PI: George Church
 Title: Targeted 2nd generation sequencing in phenotyped Framingham & PGP populations

Private Source	10/01/08 - 9/30/12
----------------	--------------------

Private Source

RC2 HG005592 (NHGRI) 9/30/09-07/31/11
 NIH-NHGRI – (Halcyon)
 PI: George Church
 Title: Development of Electron Microscopy-based Nucleic Acid Polymer Sequencing

Program Director / Principal Investigator: Church, George M.

RC2HL102815 (NHLBI) 9/30/09-08/31/11
 NIH- NHLBI
 PI: George Daley (CHB)
 Title: Comparative phenotypic, functional, and molecular analysis of ESC and iPSC

ONRBAA09-001 (ONR) 4/01/10-03/30/13
 Office of Naval Research
 PI: George Church
 Title: Multiplexed Pathway and Organism Engineering

RC1 HG005482 (NCRR) 9/22/09-06/30/11
 NIH/NCRR
 PI: Peter Park
 Title: Statistical Methods for Estimation of Copy Number from Next – Generation Sequencing

DE-AR0000079
 ARPA- E (DOE) 7/01/10-06/30/13
 Title: Engineering a Bacterial Reverse Fuel Cell
 PI: Silver, Pamela A; Co-I: George Church

CBET1033397 (NSF) 1/1/11-1/31/13
 NSF
 PI: Ryan Gill (U. Colorado)
 Title: A new approach for directed genome engineering

DARPA 11-23-CCM-DT-FP-006 6/1/11-5/31/15
 PI: Jim Collins (BU)
 Title: Synthetic Mammalian Gene Regulatory Circuits for In Vivo Biomedical Applications

ONR 6/1/11-5/31/16
 PI: Jim Collins (BU)
 Title: Utilizing Synthetic Biology to Create Programmable Micro- Bio- Robots

Completed:

P50 HG003170 (CEGS supplement) 7/1/09-06/30/10
 NIH- NHGRI
 PI: George Church
 Title: Molecular and Genomic Imaging Center

W911NF-08-1-0254 (DARPA) 6/27/08 - 1/31/11
 PI: Neil Gershenfeld
 Title: Milli-Biology: Programmed Assembly of Engineered Materials

Program Director/Principal Investigator (Last, First, Middle): Church, George M.

PROGRESS REPORT SUMMARY	GRANT NUMBER P50HG005550	
	PERIOD COVERED BY THIS REPORT	
PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR George Church	FROM 9/13/2010	THROUGH 5/31/2011
APPLICANT ORGANIZATION Harvard Medical School		
TITLE OF PROJECT (Repeat title shown in Item 1 on first page) Causal Transcriptional Consequences of Human Genetic variation		
<p>A. Human Subjects (Complete Item 6 on the Face Page)</p> <p>Involvement of Human Subjects <input checked="" type="checkbox"/> No Change Since Previous Submission <input type="checkbox"/> Change</p> <p>B. Vertebrate Animals (Complete Item 7 on the Face Page)</p> <p>Use of Vertebrate Animals <input checked="" type="checkbox"/> No Change Since Previous Submission <input type="checkbox"/> Change</p> <p>C. Select Agent Research <input checked="" type="checkbox"/> No Change Since Previous Submission <input type="checkbox"/> Change</p> <p>D. Multiple PD/PI Leadership Plan <input checked="" type="checkbox"/> No Change Since Previous Submission <input type="checkbox"/> Change</p> <p>E. Human Embryonic Stem Cell Line(s) Used <input checked="" type="checkbox"/> No Change Since Previous Submission <input type="checkbox"/> Change</p>		

SEE PHS 2590 INSTRUCTIONS.

WOMEN AND MINORITY INCLUSION: See PHS 398 Instructions. Use Inclusion Enrollment Report Format Page and, if necessary, Targeted/Planned Enrollment Format Page.

CCV progress report 2011

May 17, 2011

Coordinated activities of the Center as a whole	1
Church Lab report	2
Joung Lab report	9
Zhang Lab report	10
Daley Lab report	13

Coordinated activities of the Center as a whole

Since inauguration of the CCV-CEGS in September, 2010, our team has made significant initial progress in several Aims while also leading the effort to integrate research directions and work processes with the three other labs in the Center. During this period, eight publications and submitted manuscripts with CCV authorship have related to CEGS research (see CEGS and CEGS-supportive Publications below). The Center has a public web site and also a private wiki for coordinating the sharing of information. Approximately 7 graduate students and 15 post-docs are being trained on CEGS-related projects. Our Minority Action Plan, under direction of Lee Bitsóí, Ed.D., has hired its first post-doc, Willie Giraldo-Rosa (Ph.D. 2010, Universidad Autonoma de Madrid; MS 1998, University of Puerto Rico), and is in the process of selecting undergraduate summer interns. Beyond the four Labs comprised within it, the Center is collaborating with three other labs: The Rossi (Childrens' Hospital) and Collins (Boston University) labs are invited to monthly CCV investigators' meetings, and we are working closely with the Nilsson Lab (Uppsala, Sweden) on Aim 3 (see below).

The Center has held monthly Investigators' meetings since October, 2010 run by co-director John Aach. In early discussions, the Center opted to change its name from the original CTCHGV (Causal Transcriptional Consequences of Human Genetic Variation) to the simpler CCV (Causal Consequences of Variation). This change denotes its determination to characterize not only *transcriptionally* causative human genomic variations, but causative variations of any kind, so long as the variations cause a cellular (vs organismal) phenotype that can be detected in one of the Center's target cell lines (mainly PGP1). In essence, CCV now broadly sees its mission as the development and demonstration of efficient and scalable techniques for human reverse genetics at the cellular level.

Consequent on that decision, the Center engaged in a dual process of gathering and consolidating resources and data on its target cell lines, exploration of available cellular phenotypes, and identification of a first set of 15 genomic engineering targets. Focusing on three PGP1 cell lines (EBV-transformed B cell, fibroblast, iPS), we now have genome sequences from Complete Genomics, Inc., have just obtained RNA-seq data, and are confirming genotypes *via* 1M Infinium bead arrays (Aim 1.3). We are also generating a ~160kb PGP1 fibroblast BAC library (5x coverage) for both alteration in *E. coli* (Aim 1.1, section 5.1.1(i), MAGE-BAC/ZFN) and for experiments involving replacements of engineered long genomic segments (Aim 4.2, section 5.4.2.iii). The Zhang Lab has analyzed PGP1 and other PGP cell lines used by the CCV for mutations (6) and, using improved padlock probe technology, for allele-specific methylation and improved allele-specific expression (Aim 1.3; see Zhang Lab report). The Daley Lab has been assisting and overseeing CCV experiments on PGP1 iPS lines (Aim 2) (see Daley Lab report). The Joung Lab has recently created 37 new Zinc Finger (ZF) OPEN subsite pools (16; 17) (Aim 4.2, section 5.4.2 i.A) and is now generating ZF Nucleases (ZFNs) for 15 genomic target regions (two per target region) for the Center to effect the targeted dsDNA cuts that will both enable engineering of these regions and testing of these new pools (see Joung Lab report). They are also in the process of creating the 76 remaining subsite pools of the total of 192 needed to target any 9-mer in the genome. The 15 genomic target regions chosen by the Center will test diverse strategies, phenotypes, and challenging engineering targets, including:

- Correction of an NEK11 mutation found by the Zhang Lab in a PGP1 iPS cell line (6). NEK11 plays a role in an S-phase checkpoint and its mutation in iPS may have been selected during iPS grow-out. We will see if reversion of this mutation affects iPS growth. This research will open up the prospect of generally engineering out potentially deleterious mutations selected by iPS conversion.

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

CEGS Progress Report 2011

- 5 SNPs *cis* to 3 genes reported in the literature to cause variations in transcription (Aim 1.2). The targets are myc (23), mdm2 (2; 9), and col1a1 (8; 19). These genes were selected from a set of 11 with reported transcriptionally causative *cis* SNPs because analysis of available expression data suggests they will be expressed in at least one PGP1 cell line (including PGP1 iPS derivatives).
- 1 SNP (rs559518) *cis* to a gene that we have predicted may cause transcriptional variation. ZFN targeting of this SNP is challenging as it abuts repetitive sequence.
- HLA-B 5' and 3' regions. We aim to replace PGP1 the HLA-B coding region (2676bp) with an alternative HLA-B allele. This will provide an initial target for Aim 4.2, section 5.4.2.ii testing of the replacement of long DNA segments with ZFNs targeted at either end. HLA-B expression results in an MHC Class 1 protein at the cell surface, enabling antibody detection of the new allele as a cellular phenotype. If successful, we plan next to segmentally replace the longer HLA-B/HLA-C region (~88kbps) using ZFNs, then replace HLA-A, and generate HLA-replaced PGP1 iPS. Ultimately, these experiments will enable the engineering of donor human iPS to be histocompatible with the MHC class 1 type of any other human, providing a foundation for engineering donor human tissues so that they will be transplantable into any human recipient who needs them (see, however, (31)).
- We are also targeting a ZFN to a frankly repetitive region (telomere repeat region) with the object of exploring whether telomere disruption affects cell senescence. Careful dosing of this ZFN will be essential and we will use it to test Protein Transduction Domain (PTD; Aim 4.2, section 5.4.2.ii) or supercharged-GFP (5; 21) delivery of the ZFN protein into the cells.
- We have created PGP1 cell lines with a new integrated GFP reporter construct (in which GFP has a broken start codon) to test and compare performance of several genome engineering methods under development (e.g., see Targeted deaminases below), and are generating a ZFN to this construct.

The Church Lab will perform most of the engineering indicated above. We now turn to other Church Lab conducted research within the Center.

CEGS and CEGS-supportive Publications

1. Carr PA, Wang HH, Sterling B, Isaacs FJ, Xu G, Church GM, Jacobson JM. 2011. Enhanced Multiplex Genome Engineering through Oligonucleotide Co-selection (in revision).
2. Gore A, Li Z, Fung HL, Young JE, Agarwal S, Antosiewicz-Bourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, Lee JH, Loh YH, Manos PD, Montserrat N, Panopoulos AD, Ruiz S, Wilbert ML, Yu J, Kirkness EF, Izpisua Belmonte JC, Rossi DJ, Thomson JA, Eggan K, Daley GQ, Goldstein LS, Zhang K. 2011. Somatic coding mutations in human induced pluripotent stem cells. *Nature* 471:63-7
3. Isaacs FJ, Carr PA, Wang HH, Lajoie MJ, Sterling B, Kraal L, Tolonen AC, Gianoulis TA, Goodman D, Reppas NB, Emig CJ, Bang D, Hwang SJ, Jewett MC, Jacobson JM, Church GM. 2011. Precise Manipulation of Chromosomes *in vivo* Enables Genome-wide Codon Replacement (in press). *Science*
4. Kosuri S, Eroshenko N, Leproust EM, Super M, Way J, Li JB, Church GM. 2010. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* 28:1295-9
5. Matzas M, Stahler PF, Kefer N, Siebelt N, Boisguerin V, Leonard JT, Keller A, Stahler CF, Haberle P, Gharizadeh B, Babrzadeh F, Church GM. 2010. High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat Biotechnol* 28:1291-4
6. Mosberg JA, Lajoie MJ, Church GM. 2010. Lambda red recombineering in *Escherichia coli* occurs through a fully single-stranded intermediate. *Genetics* 186:791-9, PMID: 2975298
7. Wang HH, Xu G, Vonner A, Church GM. 2011. Modified Bases Enable High-efficiency Oligonucleotide-Mediated Allelic Replacement via Mismatch Repair Evasion (in press). *NAR*
8. Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. 2011. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotechnol* 29:149-53

Church Lab report**Aim 1: Human genome engineering techniques**

In addition to ZFNs developed by the Joung Lab, the Church Lab is developing and optimizing several human

genome engineering methods. Common goals of these non-ZFN methods include high efficiency, avoidance of the dsDNA cuts induced by ZFNs (which are toxic and induce error-prone Non-Homologous End Joining (NHEJ) DNA repair in addition to Homologous Recombination (HR)), easier targeting, and easier synthesizability. The ZFN engineering experiments above will serve as a basis for comparison for these newer techniques.

MAGE-human: In MAGE (Multiplex Automated Genome Engineering, (26)), small scale substitutions and indels, but also long deletions >10kbp, are efficiently introduced in a genome by provision of appropriately structured ssDNA oligos that are incorporated as Okazaki fragments during DNA replication. Modifications depend on λ -Red activities (29) and can be readily multiplexed. We are in the process of transferring and optimizing procedures for human cells that have been successfully used in *E. coli*, using a system in which HeLa cells containing a defective EGFP are corrected with an oligonucleotide containing the non-defective sequence (28). At this time our best results have achieved 3-5% efficiency using lipofectamine-transfected 25bp ssDNA oligos containing 12 phosphorothioate bonds. We have obtained up to ~2x improvement by suppressing mismatch repair (MMR) using shRNAs (18). However, by using chemically modified bases found to avoid MMR in *E. coli* (27), we are able to get a similar increase in efficiency without global silencing of the MMR pathway (see Figure 1a). Currently we are focusing on our observation that corrected GFP-expressing cells appear to have a lower growth rate than uncorrected cells (Figure 1b), so that corrected cells are outcompeted during grow-out. This growth rate difference is proportional to the number of phosphorothioate bonds in the oligo (Figure 1b). We have exploited this by preferentially killing the faster-dividing non-corrected cells by treating all cells with antimicrotubule chemotherapeutic drugs (Figure 1c); this reduces the decline in corrected cell frequency over time. We have compared expression profiles of corrected vs. uncorrected cells using RNA-seq and found differences which may explain the growth rate discrepancies (Figure 1d). To functionally test these differences we are planning pooled RNAi screens to find targets that decrease the oligo toxicity and increase incorporation efficiency.

To move from the model HeLa system to more experimentally relevant cell types, we have developed a similar defective-EGFP reporter line with PGP1 iPS cells and are currently characterizing this new system. We are also attempting to develop better methods of growing, maintaining, and screening of mammalian cells, particularly since single-base correction of native loci will not be amenable to fluorescent or antibiotic resistance selection.

Other MAGE: Progress in MAGE in *E. coli* has direct consequences for human

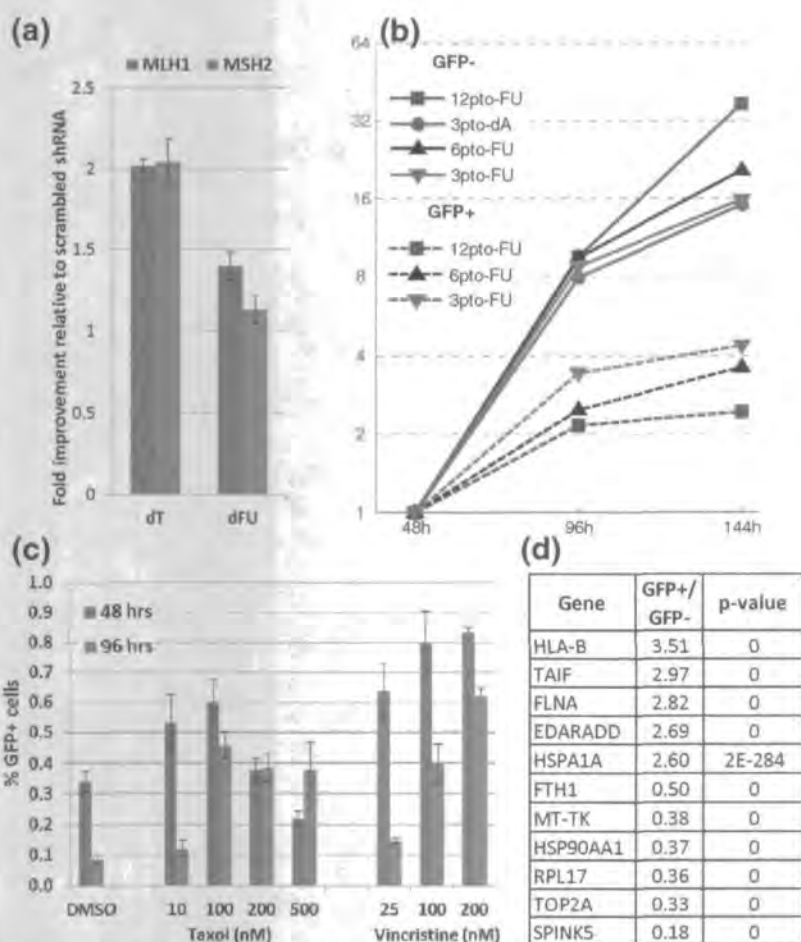


Figure 1: Oligo-mediated Human Genome Engineering. (a) RNAi against MMR components increases efficiency ~2x with normal (dT) vs modified (FU) mismatched bases. MMR suppression decreases effect of modified bases, indicating at least partial escape of modified mismatched base incorporation from MMR. (b) Proliferation of corrected vs uncorrected cells as measured with CellTrace Violet. 3pto-dA is a non-correcting control oligo. (c) Enrichment of corrected cells generated with 3pto-FU oligo by antimicrotubule drug treatments. (d) Top differentially expressed genes between corrected and uncorrected cells, determined by RNAseq.

genome engineering because targeted changes to human DNA can be made in *E. coli* and then moved to human (Aim 1.1, section 5.1.1(i) MAGE-BAC/ZFN), and because some enhancements achieved in MAGE *E. coli* can be transferred directly to MAGE-human. Along these lines, we have discovered that MAGE efficiency can be enhanced by co-selecting a defective antibiotic resistance marker that is near a targeted genomic site (3), and also by using ssDNA oligos with modified DNA bases that are not detected by mismatch repair (27). We are also attempting to increase the efficiency of modifying ~1kb regions vs short oligo-length regions using MAGE-like techniques. Supporting this, we have determined that long regions > 1kb that are incorporated into the *E. coli* genome via λ -Red-mediated recombination appear always to do so through a fully ssDNA intermediate, consistent with the Okazaki fragment pathway exploited by MAGE (22). Finally, in conjunction with our use of MAGE to replace 314 TAG stop codons with TAA codons in the *E. coli* genome, we developed an efficient method for engineering Mb sized genomic segments of DNA (CAGE = Conjugative Assembly Genome Engineering, (7)).

TAL-based genome targeting: We (30) (also see below) and others (1; 4; 14; 24) have been exploring the use of TAL DNA binding domains naturally expressed by *Xanthomonas* spp. as a possible alternative to ZF arrays for DNA targeting and modification. Unlike ZFs, TAL proteins present a simple one-to-one correspondence between the contents of diresidue positions 12-13 in a stretch of *N* concatenated 33-34 residue repeat monomers, and the corresponding single nucleotides in the *N* bases of their dsDNA binding targets. Our object in (30) was to design an efficient and modular method for assembling repeat monomers for arbitrary user-selected DNA binding sites into an optimized backbone that is compatible with high-throughput DNA construct synthesis methods we have also recently developed ((10) and see below). Features unique to our method vs recently published alternatives (15) include (i) codon adjustment of the monomer repeat units to maximize DNA sequence divergence between monomers to make for more specific PCR assembly and reduce the likelihood of *in-vivo* recombinations; and (ii) optimized reduction of the N- and C- termini of the native TAL protein backbone. We successfully generated TAL-based activators designed to bind to seventeen different 14bp sequences placed upstream of a reporter construct and tested them in human cells (293FT), finding that three achieved >50-fold induction of the reporter, all but three activated the reporter >10-fold, and all but one activated it >5-fold. We also demonstrated synthesized TAL activator reporters bearing upstream sequences taken from several human pluripotency factors (see Figure 2).

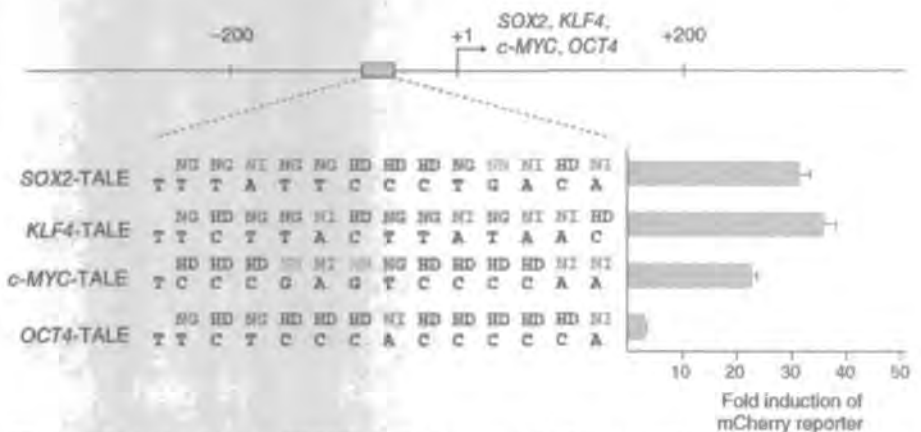


Figure 2: Induction in 293FT cells by modularly assembled TAL activators targeted to sequences upstream of Sox2, Klf4, Myc, and Oct4. From (10).

Targeted deaminases: We have been developing and optimizing DNA deaminases that are targeted to specific user-defined DNA sequences as a way of engineering specific mutations. The targeted deaminases are built as fusions between cytidine deaminases (such as Activation Induced Deaminase (AID) and members of the APOBEC family) with either ZF or TAL proteins designed to bind to specific DNA sequences. Although such deaminases can only effect transitions in the genome, they avoid the toxicity and the unpredictable mutations generated by error-prone NHEJ-repair associated with targeted nucleases such as ZFNs and TAL effector nucleases. Targeted deaminase evaluations have been conducted in an *E. coli* strain containing an integrated GFP gene with a broken start codon that can be fixed with by a C→T transition (ACG→ATG). Using this method, we have assessed the effects of the length of the linker between ZF proteins and AID, of different C-terminal truncations of the TAL backbone fused to AID, and a variety of different deaminases (see Figure 3). Off target modifications are an important consideration for targeted deaminases especially given that most of these deaminases are natively processive. Preliminary data show that TAL-AID exhibits ~2% off-target mutation rate in the vicinity of the target, while ZFP-AID exhibits a ~6% off-target mutation rate. Genome sequencing of 6 corrected and 2 control strains indicates that distal off-target editing, to the extent it occurs, is

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

CEGS Progress Report 2011

very limited in degree.

Aim 2: iPS generation and engineering

With guidance from the Daley Lab, and in collaboration with the Collins Lab, we have begun a series of experiments in which fixed, decellularized ~15.5dpc mouse embryos (which we term "embryo scaffolds") are recellularized with human iPS, to test the hypothesis that residual differentiation factors in the fixed and decellularized tissue can effectively guide niche-specific differentiation of the iPS into diverse human cell types. Preliminary results indicate that some of the scaffolds generate and exude human CD19+ B-lymphoid, myeloid, and hematopoietic stem cells.

Aim 3: In situ RNA expression

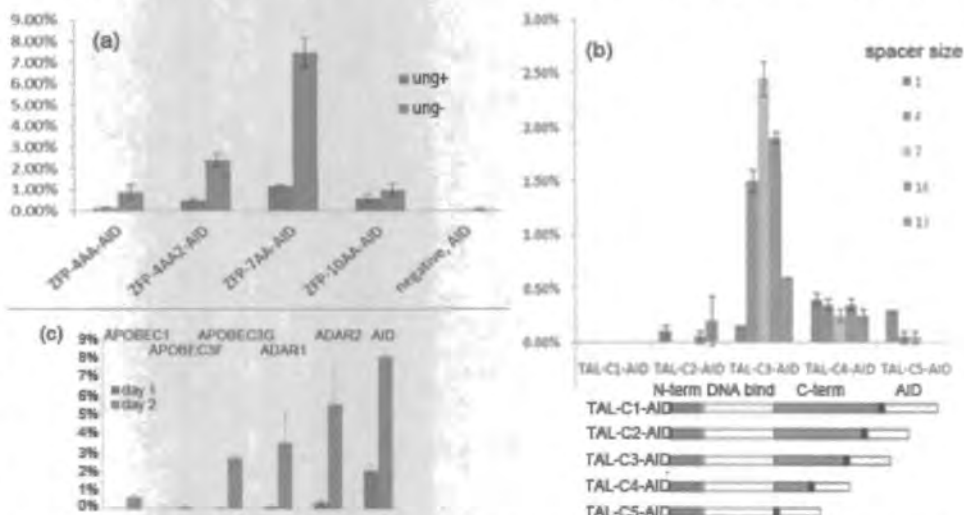


Figure 3: Performance of different targeted deaminase proteins as measured by correction of broken GFP start codon in *E. coli*. (a) ZF protein / AID fusions with different linker lengths. (b) TAL / AID fusions with different linker lengths and C-terminal TAL truncations, (c) TAL fusions with different deaminases.



Figure 4: Human and mouse β -actin transcripts in co-cultured human and mouse cells detected *in situ* by methods of (11) by Church Lab as part of collaboration with Nilsson Lab.

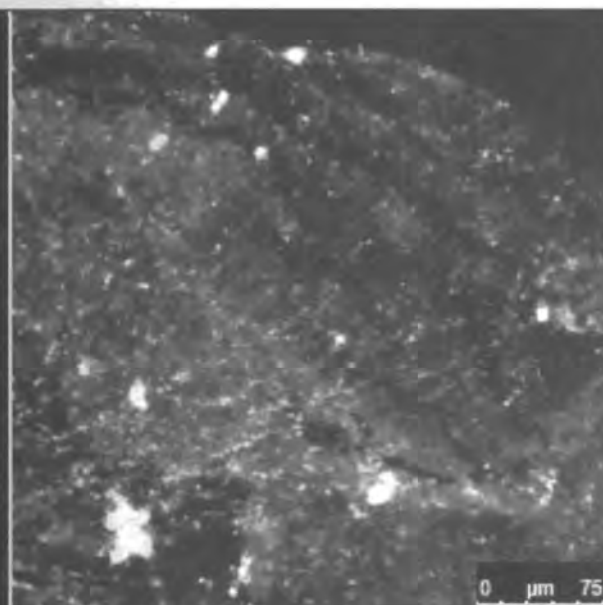


Figure 5: *In situ* Rolling Circle Amplicons (RCAs) generated from circularized FISH probes for 4000 human developmental genes in a paraffin-embedded tissue section, simultaneously sequenced at +3 position after multiple prior stripping cycles. (20x; 3-colors for the four bases (Tex-red and Cy3 combined))

In October 2010 we formed a collaboration with the Nilsson Lab (Uppsala University) with the object of increasing the multiplexity and throughput of their *in situ* mRNA expression methods (11; 12) by adapting their protocols to (i) enable their automation on the Polonator, and to (ii) use *in situ* sequencing vs hybridization probes to as a means of detecting transcripts. Nilsson Lab protocols first generate cDNA *in situ* from specific transcripts using LNA-containing primers, then circularize transcript-specific padlock probes against the cDNA, amplify the circles into Rolling Circle Amplicons (RCA) via a rolling

circle polymerase, and detect the amplicons with circle-specific probes. Transcript sequences differing at a single base have been discriminated by these methods (11). Single base pair sequencing by ligation (25) vs hybridization probing has been demonstrated to work as a method of detecting and counting transcripts. As part of the collaboration, a Church Lab graduate student spent 3 weeks in the Nilsson Lab (see Figure 4), and the Nilsson Lab is attempting to apply their methods to cells arrayed in Polonator flow cells. We are working towards assaying 10 transcripts that we have previously found to exhibit allele-specific expression in PGP1 cell lines using 20 allele-specific padlock probes, and quantifying their expression by sequencing allele-specific barcodes.

Additionally, we have been developing a FISH-seq method in which RCAs can be created *in situ* directly off of RNA without the need for *in situ* generation of single stranded cDNA. In this protocol, pre-circularized mRNA sequence-specific DNA FISH probes are introduced into permeabilized and fixed tissues where they hybridize against their target transcripts. RNase A is used to eliminate unhybridized mRNA, and phi129 polymerase and RNase III are then used to generate RCAs off of the circles using the remaining hybridized mRNA fragments as primers. RCAs of ~1 μ m diameter can be generated that can be repeatedly probed and stripped or subjected to multiple cycles of sequencing by ligation. We have been experimenting with a set of 4000 FISH probes against human developmental genes that have been divided into 40 subpools that can be individually interrogated (see Figure 5). The subpooling separates abundant transcripts so that individual pools are sparse enough that RCAs do not overlap significantly. Current priorities include reducing cell background fluorescence, and assaying quantitative accuracy and specificity.

Aim 4: Enabling technologies

High throughput DNA synthesis (Aim 4.1): We developed a high-throughput, low-error DNA synthesis pipeline in which ~13K low-error Agilent OLS oligos (130-200bp) (13) are synthesized on arrays, and then subpools of oligos for specific constructs are selectively amplified and assembled *via* PCR (10). The construct subpools (typically 10-11 oligos) are designed so that construct-specific primers are built into the Agilent oligos (see Figure 6). We tested over 1000 assembly PCR conditions and optimized rules for primer and oligo construction. Our protocols control a major difficulty in high-throughput DNA synthesis from array oligos by reducing the complexity of the oligo population in which assembly takes place from tens of thousands of oligos to just the oligos needed for a specific construct. We performed assemblies of 47 sequences (5 fluorescent proteins (FPs) and 42 antibodies; lengths 723-793bp), of which 45 produced correct length assemblies, perfect sequences were found among 18/20 cloned genes, and error rates were 1/1350bp-1/1500bp for the FPs and ~1/315bp for the (more challenging) antibodies (36). We are currently testing parallelization into 96 well plates that will be practical for synthesizing batches of ~1000 constructs and plan to substantially increase multiplexity by performing subpool amplification and assembly in emulsion.

We also report progress on the light-directed release of oligos that have been arrayed on the Polonator, another component of the integrated sequencing and synthesis concept proposed in Aim 4.1 (section 5.4.1). We have fully integrated a digital micro mirror device with an existing Polonator instrument, giving us the ability to project UV light onto a Polonator flow cell. In Polonator flow cells in which we have loaded and immobilized microbeads with synthetic DNA templates containing a photo-cleavable linker, we have achieved the following results with protocols in development: (i) We have sequenced the first base of the templates, and released templates from the beads using UV light at single pixel resolution. We have released oligos from single or multiple beads in an image field. (ii) We have collected the wash buffer used on the beads during the release process and, after rolling circle re-amplification, sequenced the amplicons and demonstrated clonality.

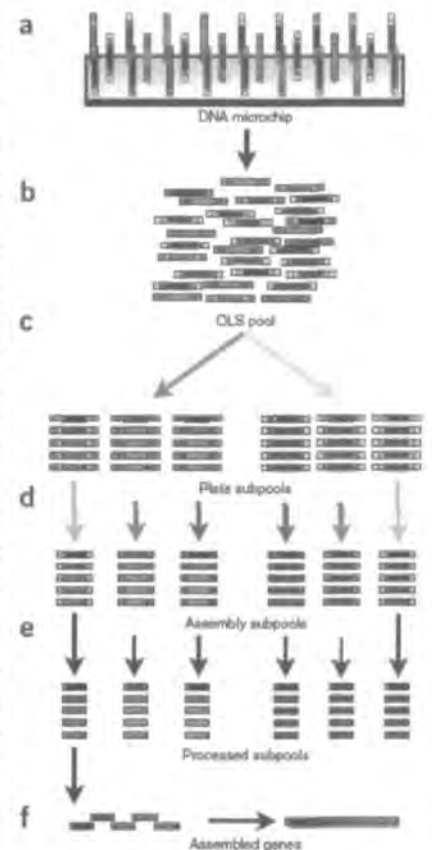


Figure 6: High-level depiction of protocol for scalable DNA construct synthesis by selective amplification and assembly of subpools of oligos generated on oligo array (10).

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

CEGS Progress Report 2011

Figure 7 illustrates the release of oligos from beads within a projected image.

In a related development, we collaborated with Febit Group (Germany) to implement an integrated sequencing and synthesis platform on the Roche 454 sequencer. In this arrangement, robotics were used to pick microbeads coated with clonally amplified oligos out of 454 picotiter plates based on their sequences (20).

References

1. Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A, Bonas U. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* 326:1509-12
2. Bond GL, Hu W, Bond EE, Robins H, Lutzker SG, Arva NC, Bargonetti J, Bartel F, Taubert H, Wuerl P, Onel K, Yip L, Hwang SJ, Strong LC, Lozano G, Levine AJ. 2004. A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell* 119:591-602
3. Carr PA, Wang HH, Sterling B, Isaacs FJ, Xu G, Church GM, Jacobson JM. 2011. Enhanced Multiplex Genome Engineering through Oligonucleotide Co-selection (in revision).
4. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF. 2010. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* 186:757-61, PMID: 2942870
5. Cronican JJ, Thompson DB, Beier KT, McNaughton BR, Cepko CL, Liu DR. 2010. Potent delivery of functional proteins into Mammalian cells in vitro and in vivo using a supercharged protein. *ACS Chem Biol* 5:747-52, PMID: 2924640
6. Gore A, Li Z, Fung HL, Young JE, Agarwal S, Antosiewicz-Bourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, Lee JH, Loh YH, Manos PD, Montserrat N, Panopoulos AD, Ruiz S, Wilbert ML, Yu J, Kirkness EF, Izpisua Belmonte JC, Rossi DJ, Thomson JA, Eggan K, Daley GQ, Goldstein LS, Zhang K. 2011. Somatic coding mutations in human induced pluripotent stem cells. *Nature* 471:63-7
7. Isaacs FJ, Carr PA, Wang HH, Lajoie MJ, Sterling B, Kraal L, Tolonen AC, Gianoulis TA, Goodman D, Reppas NB, Emig CJ, Bang D, Hwang SJ, Jewett MC, Jacobson JM, Church GM. 2011. Precise Manipulation of Chromosomes *in vivo* Enables Genome-wide Codon Replacement (in press). *Science*
8. Jin H, van't Hof RJ, Albagha OM, Ralston SH. 2009. Promoter and intron 1 polymorphisms of COL1A1 interact to regulate transcription and susceptibility to osteoporosis. *Hum Mol Genet* 18:2729-38
9. Knappskog S, Bjornslett M, Myklebust LM, Huijts PE, Vreeswijk MP, Edvardsen H, Guo Y, Zhang X, Yang M, Ylisaukko-Oja SK, Alhopuro P, Arola J, Tollenaar RA, van Asperen CJ, Seynaeve C, Staalesen V, Chrisanthar R, Lokkevik E, Salvesen HB, Evans DG, Newman WG, Lin D, Aaltonen LA, Borresen-Dale AL, Tell GS, Stoltenberg C, Romundstad P, Hveem K, Lillehaug JR, Vatten L, Devilee P, Dorum A, Lonning PE. 2011. The MDM2 promoter SNP285C/309G haplotype diminishes Sp1 transcription factor binding and reduces risk for breast and ovarian cancer in Caucasians. *Cancer Cell* 19:273-82
10. Kosuri S, Eroshenko N, Leproust EM, Super M, Way J, Li JB, Church GM. 2010. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* 28:1295-9
11. Larsson C, Grundberg I, Soderberg O, Nilsson M. 2010. In situ detection and genotyping of individual mRNA molecules. *Nat Methods* 7:395-7

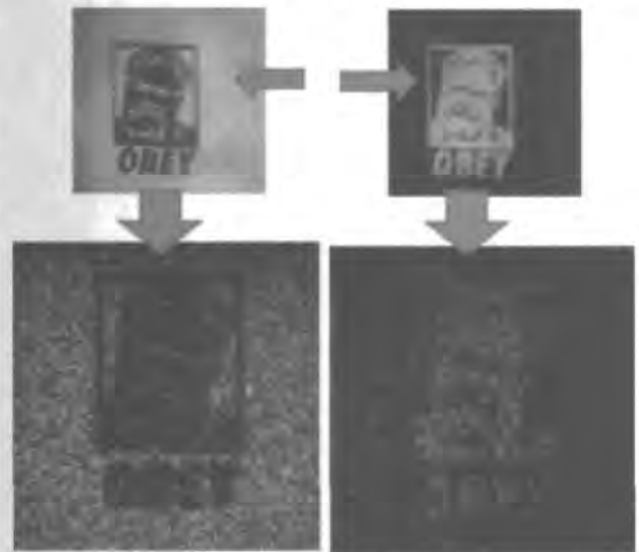


Figure 7: Illustration of light-directed release of oligos from oligo-coated beads in Polonator flow cell. Positive (top left) and negative (top right) masks displaying a human face were each projected on oligo-coated beads in a flow cell, causing release of fluorescent oligos on beads within the transparent regions of the mask (lower images)

12. Larsson C, Koch J, Nygren A, Janssen G, Raap AK, Landegren U, Nilsson M. 2004. In situ genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. *Nat Methods* 1:227-32
13. LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH. 2010. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res* 38:2522-40, PMID: 2860131
14. Li T, Huang S, Jiang WZ, Wright D, Spalding MH, Weeks DP, Yang B. 2011. TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res* 39:359-72, PMID: 3017587
15. Li T, Huang S, Zhao X, Wright DA, Carpenter S, Spalding MH, Weeks DP, Yang B. 2011. Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic Acids Res*
16. Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichinger M, Jiang T, Foley JE, Winfrey RJ, Townsend JA, Unger-Wallace E, Sander JD, Muller-Lerch F, Fu F, Pearlberg J, Gobel C, Dassie JP, Pruett-Miller SM, Porteus MH, Sgroi DC, Iafrate AJ, Dobbs D, McCray PB, Jr., Cathomen T, Voytas DF, Joung JK. 2008. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* 31:294-301, PMID: 2535758
17. Maeder ML, Thibodeau-Beganny S, Sander JD, Voytas DF, Joung JK. 2009. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat Protoc* 4:1471-501, PMID: 2858690
18. Maguire KK, Kmiec EB. 2007. Multiple roles for MSH2 in the repair of a deletion mutation directed by modified single-stranded oligonucleotides. *Gene* 386:107-14, PMID: 1847641
19. Mann V, Hobson EE, Li B, Stewart TL, Grant SF, Robins SP, Aspden RM, Ralston SH. 2001. A COL1A1 Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. *J Clin Invest* 107:899-907, PMID: 199568
20. Matzas M, Stahler PF, Kefer N, Siebelt N, Boisguerin V, Leonard JT, Keller A, Stahler CF, Haberle P, Gharizadeh B, Babrzadeh F, Church GM. 2010. High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat Biotechnol* 28:1291-4
21. McNaughton BR, Cronican JJ, Thompson DB, Liu DR. 2009. Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins. *Proc Natl Acad Sci U S A* 106:6111-6, PMID: 2659711
22. Mosberg JA, Lajoie MJ, Church GM. 2010. Lambda red recombineering in Escherichia coli occurs through a fully single-stranded intermediate. *Genetics* 186:791-9, PMID: 2975298
23. Pomerantz MM, Ahmadiyah N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, Yao K, Kehoe SM, Lenz HJ, Haiman CA, Yan C, Henderson BE, Frenkel B, Barretina J, Bass A, Tabernero J, Baselga J, Regan MM, Manak JR, Shivdasani R, Coetzee GA, Freedman ML. 2009. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* 41:882-4, PMID: 2763485
24. Scholze H, Boch J. 2010. TAL effector-DNA specificity. *Virulence* 1:428-32
25. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728-32
26. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM. 2009. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460:894-8
27. Wang HH, Xu G, Vonner A, Church GM. 2011. Modified Bases Enable High-efficiency Oligonucleotide-Mediated Allelic Replacement via Mismatch Repair Evasion (in press). *NAR*
28. Yin WX, Wu XS, Liu G, Li ZH, Watt RM, Huang JD, Liu DP, Liang CC. 2005. Targeted correction of a chromosomal point mutation by modified single-stranded oligonucleotides in a GFP recovery system. *Biochem Biophys Res Commun* 334:1032-41
29. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL. 2000. An efficient recombination system for chromosome engineering in Escherichia coli. *Proc Natl Acad Sci U S A* 97:5978-83, PMID: 18544
30. Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. 2011. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat Biotechnol* 29:149-53
31. Zhao T, Zhang Z-N, Rong Z, Xu Y. 2011. Immunogenicity of induced pluripotent stem cells. *Nature*

advance online publication

Joung Lab report

In the first eight months of this CEGS grant, the Joung lab has made progress in the following areas of Aim 4.2:

(A) Progress toward engineering a comprehensive zinc-finger archive.

(i) Construction of new OPEN pools for additional three bp target subsites

The Joung lab originally described a method termed Oligomerized Pool Engineering (OPEN) for constructing three-finger arrays (Maeder et al., Mol Cell 2008). With the OPEN method, pools of zinc fingers for different 3 bp subsites are recombined together to create a combinatorial library specific for a given 9 bp target site. Bacterial cell-based selections are then used to identify members of this library that bind efficiently to the target site of interest.

Fully enabling OPEN to target any 9 bp site of interest will require zinc finger pools for 192 different subsites (64 subsites at each of the three finger positions; Table 1). The goal of obtaining all 192 pools is also important for implementation of the strategy to obtain a comprehensive archive of zinc finger arrays for all possible 9 bp target sites proposed in our CEGS grant. The Joung lab originally described a set of zinc finger pools targeted to 66 different three bp subsites (Maeder et al., Mol Cell 2008; Table 1, grey colored boxes) and then subsequently constructed pools for an additional 13 bp subsites (Table 1, blue colored boxes).

F1				F2				F3			
GAA	GCA	GGA	GTA	GAA	GCA	GGA	GTA	GAA	GCA	GGA	GTA
GAC	GCC	GGC	GTC	GAC	GCC	GGC	GTC	GAC	GCC	GGC	GTC
GAG	GCG	GGG	GTG	GAG	GCG	GGG	GTG	GAG	GCG	GGG	GTG
GAT	GCT	GGT	GTT	GAT	GCT	GGT	GTT	GAT	GCT	GGT	GTT
TAA	TCA	TGA	TTA	TAA	TCA	TGA	TTA	TAA	TCA	TGA	TTA
TAC	TCC	TGC	TTC	TAC	TCC	TGC	TTC	TAC	TCC	TGC	TTC
TAG	TCG	TGG	TTG	TAG	TCG	TGG	TTG	TAG	TCG	TGG	TTG
TAT	TCT	TGT	TTT	TAT	TCT	TGT	TTT	TAT	TCT	TGT	TTT
AAA	ACA	AGA	ATA	AAA	ACA	AGA	ATA	AAA	ACA	AGA	ATA
AAC	ACC	AGC	ATC	AAC	ACC	AGC	ATC	AAC	ACC	AGC	ATC
AAG	ACG	AGG	ATG	AAG	ACG	AGG	ATG	AAG	ACG	AGG	ATG
AAT	ACT	AGT	ATT	AAT	ACT	AGT	ATT	AAT	ACT	AGT	ATT
CAA	CCA	CGA	CTA	CAA	CCA	CGA	CTA	CAA	CCA	CGA	CTA
CAC	CCC	CGC	CTC	CAC	CCC	CGC	CTC	CAC	CCC	CGC	CTC
CAG	CCG	CGG	CTG	CAG	CCG	CGG	CTG	CAG	CCG	CGG	CTG
CAT	CCT	CGT	CTT	CAT	CCT	CGT	CTT	CAT	CCT	CGT	CTT

Table 1 Summary of OPEN pools constructed. Target subsites for all 192 potential pools are listed -- 64 different three bp subsites for each finger position (F1 = finger 1, F2 = finger 2, and F3 = finger 3). Pools originally described in Maeder et al., Mol Cell 2008 are colored in grey. Additional unpublished pools from the Joung lab previously described in the CEGS application are colored in blue. New pools obtained since the start of the CEGS grant are colored in green. Remaining uncolored subsites are in progress.

In the last eight months, the Joung lab has initiated selections for all 113 remaining three bp subsites. To date, we have obtained pools for 37 additional subsites (Table 1, green colored boxes). Sequencing of a small number of fingers from each of these pools reveals collections of fingers that resemble one another but still show amino acid diversity (as was the case for the original 79 pools). In total, the Joung lab now has zinc finger pools for 116 of the 192 possible subsites. In the coming year, we plan to continue and complete selection of finger pools for the remaining 76 subsites.

ii. Validation of new OPEN pools and selection of ZFNs for endogenous gene targets

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

CEGS Progress Report 2011

To further validate the 37 additional pools described above, the Joung lab is currently undertaking a set of OPEN selections for 90 different 9 bp sites that use the new pools. The sequences of these 90 sites are shown in Table 2. In addition to validating the new pools, these particular selections were chosen so that the resulting zinc finger arrays could be used to create ZFN pairs for 45 target sites in various mammalian genes. These ZFNs will be used by various members of the Church group for other projects that are part of the CEGS grant.

(B) Progress toward development of a segmental replacement strategy

The CEGS proposal described the development of an approach to replace segments of DNA located between two double-strand breaks introduced by two pairs of ZFNs. The Joung lab, working with the lab of Toni Cathomen at Hannover Medical School, also recently described orthologous pairs of FokI domain heterodimers. These domains can be used to express two pairs of heterodimeric ZFNs with minimal cross-talk between the pairs (i.e.--minimal formation of undesired heterodimers) (Söllü et al., Nucleic Acids Res. 2011). Using these orthologous domains, we are co-

Target gene name	Target half-site (5' to 3')	Target gene name	Target half-site (5' to 3')
mGFP_L	GTGGCCGTA	REF_PGP1_PGP1_NEK11_L	GCAAAAGCA
mGFP_R	CGCGGGGTT	REF_PGP1_PGP1_NEK11_R	AGTGTAGTG
mGFP_L	GTAGGTCTG	REF_PGP1_PGP1_NEK11_L	GCAAAAGCA
mGFP_R	CACGCGGGG	REF_PGP1_PGP1_NEK11_R	GTAGTGTAG
mGFP_L	GCGGCAGAA	REF_PGP1_PGP1_NEK11_L	GGAGCTTCC
mGFP_R	CGATGGGGA	REF_PGP1_PGP1_NEK11_R	GCTGATGAG
HLA_L	GCTGTCTCA	REF_RS6983267_L	TCTGCTGAG
HLA_R	GTGAGGGAC	REF_RS6983267_R	GGCACTGAG
HLA_L	ACAGCTGTC	REF_RS6983267_L	GATAGGGAG
HLA_R	GGGACTGAG	REF_RS6983267_R	ACAGAGGGA
HLA_L	GGTGCGTGG	RS2412298_RS1107946_L	GGGGGCAGA
HLA_R	GACTCAGAG	RS2412298_RS1107946_R	GTGCCAGCG
HLA_L	GACTCTGAG	RS2412298_RS1107946_L	CTAGGCGGG
HLA_R	GACGCGGAG	RS2412298_RS1107946_R	GCGACTGCA
HLA_L	GTGCGTGGG	RS2412298_RS1107946_L	AGGGAGGGG
HLA_R	GACTCAGAG	RS2412298_RS1107946_R	GCGACAGGG
IGH_L	GAAATAGGG	RS2412298_RS1107946_L	GGGGCAGAC
IGH_R	GCATCGGAA	RS2412298_RS1107946_R	GTGGCAGCG
IGH_L	AGAGCAGGG	RS2412298_RS1107946_L	TAGGCGGGG
IGH_R	ACCGGGGCT	RS2412298_RS1107946_R	GCGACTGCA
IGH_L	GTAAGTGGG	RS2412298_RS1107946_L	TAGGAGGGG
IGH_R	GACGCACCC	RS2412298_RS1107946_R	GACAGGGGT
IGK_L	GTAAGAGAG	RS2412298_RS1107946_L	GGAGGGGGT
IGK_R	AAATAGGTA	RS2412298_RS1107946_R	GCGACAGGG
IGK_L	GAAGAGGCA	RS2412298_RS1107946_L	AGGGAGGGG
IGK_R	GCAATAGG	RS2412298_RS1107946_R	GACAGGGGT
LIMD_A (MOUSE)_L	GGTCCATCC	RS2412298_RS1107946_L	GCTCTGGAG
LIMD_A (MOUSE)_R	GAGACTGCC	RS2412298_RS1107946_R	GTGGAAGCG
LIMD_A (MOUSE)_L	TCCATCCTT	RS559518_L	GAGGAACCT
LIMD_A (MOUSE)_R	GAGACTGCC	RS559518_R	TGTGCTGCA
LIMD_F3 (MOUSE)_L	AGTCAAGAG	RS559518_L	TCAGAGGTC
LIMD_F3 (MOUSE)_R	GCATATGTA	RS559518_R	GTTGTACAC
LIMD_F3 (MOUSE)_L	GTCAGAGGC	RS7484572_RS2279744_L	GCCGCAGCG
LIMD_F3 (MOUSE)_R	GCATATGTA	RS7484572_RS2279744_R	GGTCCGGAT
MM_CDKN2A_DOWNSTREAM_L	GATATAGAA	RS7484572_RS2279744_L	GCCGCAGCG
MM_CDKN2A_DOWNSTREAM_R	ACAGAAGGT	RS7484572_RS2279744_R	AGGTCCGGA
MM_CDKN2A_DOWNSTREAM_L	GTGTCTGTG	RS7484572_RS2279744_L	GTGTCTGAA
MM_CDKN2A_DOWNSTREAM_R	GAAATAGTT	RS7484572_RS2279744_R	GAAACTGCA
MM_CDKN2A_DOWNSTREAM_L	GATGTGTC	RS7484572_RS2279744_L	GCACAAAGG
MM_CDKN2A_DOWNSTREAM_R	GGCACAGTG	RS7484572_RS2279744_R	GTGGCTGGG
MM_CDKN2B_UPSTREAM_L	GAGGACTCA	telomere repeat_L	CTAACCTTA
MM_CDKN2B_UPSTREAM_R	GACCAAGGT	telomere repeat_R	GGTTAGGGT
MM_CDKN2B_UPSTREAM_L	GTGTGTGGC	telomere repeat_L	CTAACCTTA
MM_CDKN2B_UPSTREAM_R	AGTGTAGAG	telomere repeat_R	GTTAGGGTT
MM_CDKN2B_UPSTREAM_L	GTGTATGAG		
MM_CDKN2B_UPSTREAM_R	GTCAAAGAT		

expressing two pairs of heterodimeric ZFNs targeted to sites in the human HoxB13 gene (HX587 and HX761) in 293 cells with the goal of replacing the ~150 bps of sequence flanked by these sites using homologous recombination. For these experiments we are using a donor template designed to re-code the segment of DNA between the two ZFN sites in a translationally silent fashion and to introduce a restriction site.

In the coming year, if we can optimize efficient segmental replacement of this small section of the HoxB13 gene (as judged by a combination of PCR-based assays, Southern blotting, and direct allele sequencing), we will attempt to perform additional replacements with larger lengths of DNA. In addition, we will begin to test the efficiency of these strategies in human induced pluripotent stem cells. For both of these applications, BAC-based donor templates being may be useful for increasing the efficiencies of such recombinations. A collection of BAC clones being derived by the Church group from PGP cells will be very important for constructing these longer length donor templates.

Table 2. List of sites targeted by OPEN selections for validation of new finger pools. Additional details in the text.

Zhang Lag CCV Progress Report 2011

Improvement of ASE assay. We previously designed two sets of padlock probes for detecting allele-specific gene expression (ASE) in the human genome, and demonstrated that ASE is both common, tissue specific, and a strong indicator for the presence of cis-regulatory variants (Lee et al. 2009; Zhang et al. 2009). There were two remaining technical limitations: the capture efficiencies were still far from even, and the presence of allele-specific capture bias. We recently designed and synthesized a third generation of probe set with significant improvements on both aspects. We used a neural network model that was trained with experimental data to design probes with a lower variation in capture efficiency. In addition, the gap size of the padlock capturing arms was increased to 18nt, such that the variant sites locate outside the footprint of DNA ligase and away from the extension terminal of DNA polymerase. Finally, we also increased the size of probe set from 27,000 to 36,000 coding SNPs. With 6-8 million Illumina 36bp sequencing reads, we were able to confidently call ~30,000 SNPs on genomic DNA and 22,000 expressed SNPs on cDNA, among which roughly 5,000 are heterozygous informative SNPs. We can multiplex at least 30 samples in one Illumina GA IIx sequencing flowcell, and 100 samples in HiSeq sequencing flowcell, making it a very efficient and cost effective assay for ASE. We are in the process of performing a more complete survey of ASE in the three cell types (EBV-transformed lymphocytes, primary fibroblasts, induced pluripotent stem cells) derived from PGP1, by combining padlock capture with this third generation probe set with high-coverage RNAseq.

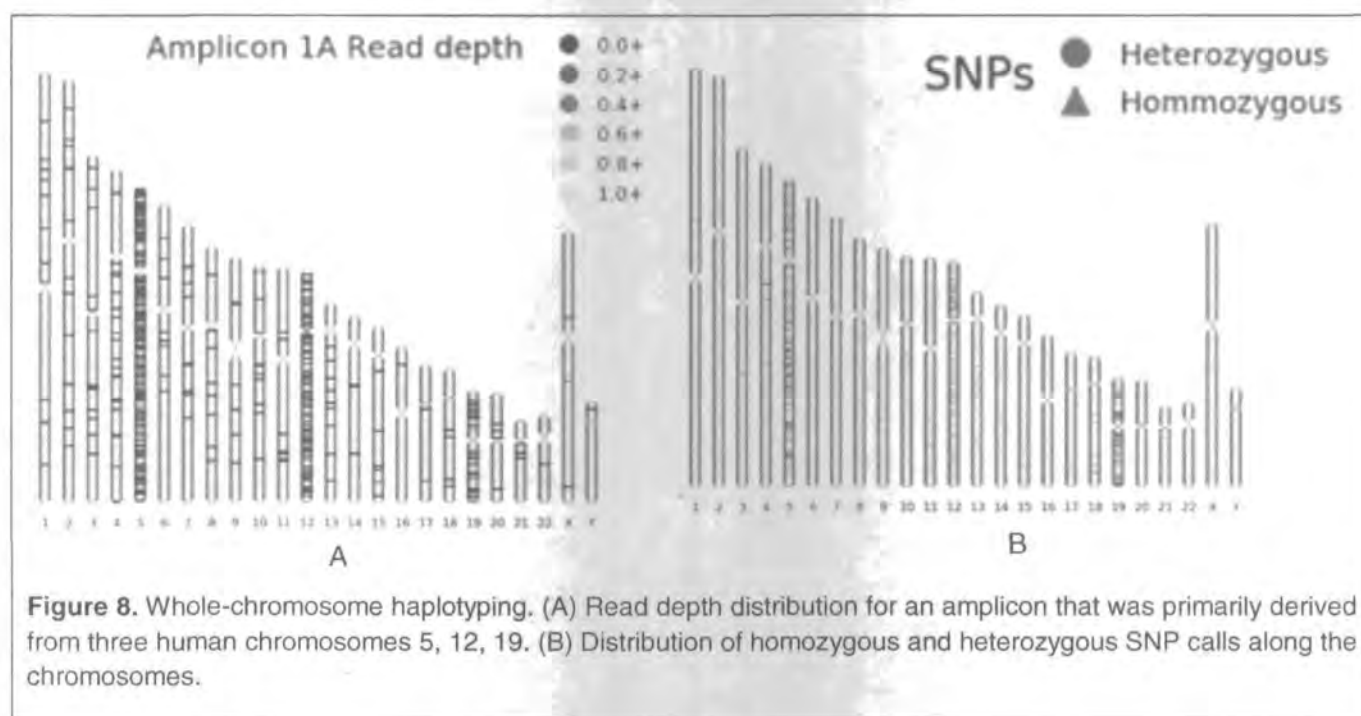
Allele-specific methylation (ASM) analysis. We have recently discovered that the presence of cis-regulatory genetic variants not only leads to allele-specific gene expression (ASE), it also results in differential epigenetic modifications on the two parental chromosomes, such as allele-specific methylation (ASM). In a recent study focusing on all CpG islands on two human chromosomes, we showed that ASM is present in the human genome at a similar level as ASE, and one mechanism through which genetic polymorphisms interact with the epigenome is direct creation or removal of CpG dinucleotides by genetic variants (Shoemaker et al. 2010). More recently, we have further expanded our survey of ASM to all 23 human chromosomes. Using a library of 330,000 bisulfite padlock probes, we quantified over 500,000 CpG sites in known differentially methylated regions and DNase I hypersensitive, and discovered thousands of ASM regions in the three PGP1 cell lines as well as additional EBV-transformed lymphocytes derived from other PGP donors and HapMap individuals (Table 3). We will integrate ASM and ASE events identified in the same set of samples to further understand how cis-regulatory variants regulate both the epigenome and transcriptome.

Whole genome haplotyping. We proposed to connect cis-regulatory genetic variants with their regulating genes by building complete diploid genome maps in which all genetic variants on 23 chromosomes are fully phased. Our strategy for generating a fully phased diploid genome involved preparation of metaphase chromosome molecules, followed by whole genome amplification of random subsets of human chromosomes, and shotgun sequencing. We have gone through one complete cycle of this experiment, and identified two sequencing libraries that were derived from a small number of human chromosomes. As shown in Figure 8, these results are promising in that single intact human chromosomes can be prepared from metaphase trapped cells and be amplified relatively even across the chromosomes; in addition, variants called on the three intact chromosomes are mostly homozygous, indicating that only one particular parental allele was present. There are two remaining issues: (i) the presence of sequencing reads from other random chromosomal locations; (ii) relative sparse sequencing coverage on the intact chromosomes, which is similar to what was recently reported with microfluidic-base MDA (Fan et al. 2011). The first issue is mostly likely due to the presence of small chromosomal fragments in the amplification templates. A simple solution is to use FACS sorting to separate intact chromosomes from small DNA fragments that were generated as the side-products of chromosome preparation. We are in the process of optimizing protocols for chromosome sorting. To improve the sequencing coverage, we are tested three different strategies, including (i) tuning down the yield of whole genome amplification (to reduce amplification bias) coupled with developing library-construction protocols that require low-input DNA; (ii) increasing the number of amplicons for each sample in order for upsampling of chromosomes; (iii) significantly increase the amount of sequencing per sample. We are optimistic that the diploid genome of PGP1 will be completely phased within one year since the start of this project (by October 2011).

Table 3

Sample	# raw reads	#CpGs called	#LD blocks	#SNPs called	# het SNPs called	#ASM I ⁺	#ASM II ⁺	#ASM III ⁺
PGP1L	22,134,386	461,908	12,437	17,240	8,787	1,652	167	1,042
PGP1F*	11,894,560	204,582	4,253	8,651	4,717	768	94	507
PGP1iPS*	14,982,406	201,255	4,589	8,641	4,455	407	123	911
PGP2L	28,495,192	543,140	11,153	19,767	10,116	1,131	191	1,545
PGP3L	29,570,157	521,279	15,896	19,235	9,884	1,757	198	1,205
PGP5L	11,576,285	308,050	7,588	12,355	6,350	643	92	882
PGP7L	11,970,974	316,178	8,347	11,917	6,138	938	130	704
PGP8L	26,422,937	484,298	12,069	17,038	8,429	1,359	210	1,208
PGP9L	29,867,962	523,066	11,602	19,164	10,057	979	181	1,660
PGP10L	33,157,508	554,246	13,504	22,369	12,045	2,033	243	1,714
GM12878	26,669,930	507,086	11,765	18,388	9,873	1,948	198	953

Notes: * PGP1F and PGP1iPS were captured with an earlier version of 220k probe set, therefore fewer CpG sites were analyzed. +Allele Specific Methylation (ASM) was divided into three classes: (I) ASM regions that are completely independent of the presence of CpG-SNPs; (II) ASM regions that contain CpG SNPs in the dbsnp131 database, but were not called heterozygous; (III) ASM regions that are due to the presence of heterozygous CpG-SNPs



References:

- Fan HC, Wang J, Potanina A, Quake SR. 2011. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* **29**(1): 51-57.
- Lee JH, Park IH, Gao Y, Li JB, Li Z, Daley GQ, Zhang K, Church GM. 2009. A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet* **5**(11): e1000718.
- Shoemaker R, Deng J, Wang W, Zhang K. 2010. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.*

Program Director/Principal Investigator (Last, First, Middle): Church, George, M.

CEGS Progress Report 2011

Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, Leproust EM et al. 2009. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* 6(8): 613-618.

Daley Lab CCV Progress Report 2011

We are providing technical support for the derivation, culture and characterization of induced pluripotent stem cells (iPSC) for the PGP cell lines used in the CCV (1); in particular linking the analysis of allele-specific regulatory variation to *in vitro* differentiation of iPSC, and performing *in vitro* differentiation to analyze epigenetic memory in iPSC that might reflect regulatory features of the tissue of origin. We are also collaborating on a novel platform for studying differentiation of iPSC into multiple tissues *via* colonization of fixed and decellularized embryo scaffolds (see Church Lab report). Hematopoietic elements are the most numerous cell populations to be detected thus far in such cultures, and preliminary analysis suggests liberation of CD19+ B lymphoid populations. Future efforts will be aimed at demonstrating the behavior of injected populations of iPSC in the embryonic matrix relevant to migration and local differentiation. Dr. Daley is also providing guidance on the ethical oversight and evaluation of the embryo scaffold project.

References

1. Gore A, Li Z, Fung HL, Young JE, Agarwal S, Antosiewicz-Bourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, Lee JH, Loh YH, Manos PD, Montserrat N, Panopoulos AD, Ruiz S, Wilbert ML, Yu J, Kirkness EF, Izpisua Belmonte JC, Rossi DJ, Thomson JA, Eggan K, Daley GQ, Goldstein LS, Zhang K. 2011. Somatic coding mutations in human induced pluripotent stem cells. *Nature* 471:63-7

2011 CCV Minority Action Plan Progress Report

Dr. Lee Bitsoi, MAP Program Director, has accomplished the following goals and objectives:

1. Organized and implemented the CCV minority student recruitment plan as outlined in the MAP proposal.
 2. Recruited and hired one Post-Doctoral Fellow:
 - a. We have selected Willie Giraldo-Rosa, Ph.D. to serve as our post-doctoral fellow this year. Dr. Giraldo-Rosa earned his B.S. (1994) and M.S. (1998) from the University of Puerto Rico and received his Ph.D. (2010) from the Universidad Autonoma de Madrid.
 - b. We intend to hire another post-doctoral by the end of this year.
 - c. Yemi Adesokan, prior MAP Post-Doctoral Fellow successfully completed his fellowship in March 2011 and has established his own start-up company—Pathogenica.
 3. Recruited the following students to participate in the CCV summer undergraduate research program this year:
 - a. Personal Info Junior, Grinnell College
 - b. Personal Info Junior, Boston College
 - c. Personal Info Junior, University of North Texas
 - d. Personal Info Sophomore, Dillard University
 - e. Personal Info Junior, Stanford University
 - f. Personal Info Junior, Trinity College
- NOTE: The following arrangements have been made for all summer interns:
- i. Matched all students with appropriate mentors.
 - ii. Arranged to have summer research interns to be housed on the Northeastern University Campus.
 - iii. Arranged to have summer research interns to take the GRE Test Prep Course to be hosted by the Broad Institute.
4. Attended all required NHGRI MAP meetings and participated in telephone conference calls to:
 - a. Establish the DACC/Redcap database for all previous MAP participants at the Washington University in St. Louis.
 - b. Continue to collect and enter data.
 - c. Successfully completed the CITI Bioethics Research Certification to allow for access to DACC/Redcap database and other IRB related matters.
 5. Worked with Yveta Masarova and John Aach to project next year's MAP budget.
 6. Recruited at minority conferences such as the Society for Advancement of Chicanos/Latinos in Science, American Indian Science and Engineering Society by conducting information sessions for prospective students.
 7. Partnered with Harvard University programs and offices to provide a support network for program participants.
 8. Collaborate with Boston Area MAP (BAMAP) centers/programs to:
 - Exchange applicants from prospective students.
 - Discuss annual NHGRI meetings.
 - Provide housing options.

Program Director/Principal Investigator (Last, first, middle): Church, George M.

GRANT NUMBER:

P50HG005550

CHECKLIST**1. PROGRAM INCOME (See instructions.)**

All applications must indicate whether program income is anticipated during the period(s) for which grant support is requested. If program income is anticipated, use the format below to reflect the amount and source(s).

Budget Period	Anticipated Amount	Source(s)

2. ASSURANCES/CERTIFICATIONS (See instructions.)

In signing the application Face Page, the authorized organizational representative agrees to comply with the policies, assurances and/or certifications listed in the application instructions when applicable. Descriptions of individual assurances/certifications are provided in Part III of the PHS 398, and listed in Part I, 4.1 under Item 14. If unable to certify compliance, where applicable, provide an explanation and place it after the Progress Report (Form Page 5).

3. FACILITIES AND ADMINISTRATIVE (F&A) COSTS

Indicate the applicant organization's most recent F&A cost rate established with the appropriate DHHS Regional Office, or, in the case of for-profit organizations, the rate established with the appropriate PHS Agency Cost Advisory Office.

F&A costs will **not** be paid on construction grants, grants to Federal organizations, grants to individuals, and conference grants. Follow any additional instructions provided for Research Career Awards, Institutional National Research Service Awards, Small Business Innovation Research/Small Business Technology Transfer Grants, foreign grants, and specialized grant applications.

☒ DHHS Agreement dated: 1/13/2011 ☐ No Facilities and Administrative Costs Requested.

☐ No DHHS Agreement, but rate established with _____ Date _____

CALCULATION*

Entire proposed budget period: Amount of base \$ 1,812,696 x Rate applied 69 % = F&A costs \$ 1,250,760

Add to total direct costs from Form Page 2 and enter new total on Face Page, Item 8b.

*Check appropriate box(es):

☐ Salary and wages base ☒ Modified total direct cost base ☐ Other base (Explain)

☐ Off-site, other special rate, or more than one rate involved (Explain)

Explanation (Attach separate sheet, if necessary.):

Harvard F&A rate: 69%

Program Director/Principal Investigator (Last, First, Middle): CHurch, George M.

ALL PERSONNEL REPORT

GRANT NUMBER

P50HG005550

Place this form at the end of the signed original copy of the application. Do not duplicate.

Always list the PD/PI(s). In addition, list all other personnel who participated in the project during the current budget period for at least one person month or more, regardless of the source of compensation (a person month equals approximately 160 hours or 8.3% of annualized effort). Use the following abbreviated categories for describing Role on Project:

- PD/PI
- Co-Investigator
- Faculty Collaborator
- Staff Scientist (doctoral level)
- Postdoc (Postdoctoral Scholar, Fellow, or Other Postdoctoral Position)
- Grad Rsch Asst (Graduate Research Assistant)
- Undergrad Rsch Asst (Undergraduate Research Assistant)
- Rsch Asst (Research Assistant/Coordinator)
- Technician
- Consultant
- Biostatistician
- Other (Specify)

If personnel are supported by a Reentry or Diversity Supplement or American Recovery and Reinvestment Act (ARRA) funding, please indicate such after the Role on Project, using the following abbreviations: RS - Reentry Supplement; DS - Diversity Supplement; AF - General ARRA Supplement; ASE - ARRA Summer Experience funding.

Use Cal (calendar), Acad, or Summer to enter months devoted to project.

Commons ID	Name	Degree(s)	SSN (last 4 digits)	Role on Project	DoB (MM /YY)	Cal	Acad	Summer
eRA Commons User Name	George Church	PhD	Personal Info	PD/PI	Personal Info	EFFORT		
	John Aach	PhD		Senior Scientist				
	Yveta Masarova	MS		Coordinator				
	Sara Vassallo	BA		Lab Tech				
	Rich Terry	MS		Engineer				
	Francois Vigneault	PhD		Postdoc				
	Jehyuk Lee	PhD		PostDoc				
	Michael Sismour	PhD		Postdoc				
	Le Cong	BS		Grad St				
	Sasha Wait Zaranek	PhD		Postdoc				
	Madeleine Price Ball	PhD		Postdoc				
	Adrian Briggs	PhD		Postdoc				
	Dan Mandel	PhD		Postdoc				

Program Director/Principal Investigator (Last, First, Middle): CHurch, George M.

ALL PERSONNEL REPORT

GRANT NUMBER

P50HG005550

Place this form at the end of the signed original copy of the application. Do not duplicate.

Always list the PD/PI(s). In addition, list all other personnel who participated in the project during the current budget period for at least one person month or more, regardless of the source of compensation (a person month equals approximately 160 hours or 8.3% of annualized effort). Use the following abbreviated categories for describing Role on Project:

- PD/PI
- Co-Investigator
- Faculty Collaborator
- Staff Scientist (doctoral level)
- Postdoc (Postdoctoral Scholar, Fellow, or Other Postdoctoral Position)
- Grad Rsch Asst (Graduate Research Assistant)
- Undergrad Rsch Asst (Undergraduate Research Assistant)
- Rsch Asst (Research Assistant/Coordinator)
- Technician
- Consultant
- Biostatistician
- Other (Specify)

If personnel are supported by a Reentry or Diversity Supplement or American Recovery and Reinvestment Act (ARRA) funding, please indicate such after the Role on Project, using the following abbreviations: RS - Reentry Supplement; DS - Diversity Supplement; AF - General ARRA Supplement; ASE - ARRA Summer Experience funding.

Use Cal (calendar), Acad, or Summer to enter months devoted to project.

Commons ID	Name	Degree(s)	SSN (last 4 digits)	Role on Project	DoB (MM /YY)	Cal	Acad	Summer
eRA Commons User Name	Hamid Mukhtar	PhD	Personal Info	Fulbright Scholar	Personal Info	EFFORT		
	Jong Kim	PhD		Postdoc				
	Volker Busskamp	PhD		Postdoc				
	Chris Gregg	PhD		Postdoc				
	Prashant Mali	PhD		Postdoc				
	Xavier Rios	BS		Grad St				
	Kim Robasky	BS		Grad St				
	Joyce Yang	BA		Grad St.				
	Luhan Yang	BS		Grad St.				
	Uri Laserson	BA		Grad St.				
	Will Giraldo-Rosa	PhD		Postdoc				
	Lee Bitsol	Ed.D		MAP coordinator				

PHS 398/2590 OTHER SUPPORT

Provide active support for all key personnel. Other Support includes all financial resources, whether Federal, non-Federal, commercial or institutional, available in direct support of an individual's research endeavors, including but not limited to research grants, cooperative agreements, contracts, and/or institutional awards. Training awards, prizes, or gifts do not need to be included.

GEORGE M. CHURCH, PhD.**ACTIVE:****DE-FG02-02ER63445 (GTL)**

2/1/2003 – 11/30/2011

EFFORT

DOE-GTL

\$1,235,477

PI: George Church

Title: Microbial Ecology, Proteogenomics & Computational Optima.

Project studies proteomics and cell models for Prochlorococcus and Pseudomonas

1P50 HG005550 (CEGS)

NIH- NHGRI

9/13/10-7/31/15

EFFORT

PI: George Church

Church: \$2,400,313 SUBS: \$ 553,889 TC/yr

Title: "Center for Causal Transcriptional Consequences of Human Genetic Variation"

Role: The Center for Transcriptional Consequences of Human Genetic Variation (CTCHGV) will develop innovative and powerful genetic engineering methods and use them to identify genetic variations that causally control gene transcription levels.

SA5283-11210 (NSF)

7/1/2010 – 6/30/2014

EFFORT

NSF-(SynBERC)

Sub: Church: \$130,596

PI: Jay Keasling (UC Berkeley)

Title: Synthetic Biology Engineering Research Center

Project role is to develop synthetic bacterial genome "chasses" for safe use in mammals

RO1 HL 094963 (NHLBI)

9/30/2008-6/30/2011 (NCE)

EFFORT

NIH - NHLBI

Church: \$369,700

PI: George Church

Title: Targeted 2nd generation sequencing

in phenotyped Framingham & PGP populations.

Private Source

10/01/08 - 9/30/12

EFFORT

Church: \$150,000

PI: George Church

Private Source

Project: The identification and characterization of naked mole-rat genes that contributed to the evolution of a long lifespan in this species.

RC2 HG005592 (NHGRI)

9/30/09-7/31/11

EFFORT

NIH-NHGRI - Halcyon

Church: \$109,958

PI: George Church

Sub: Halcyon: 1,000,000/yr

Title: Development of Electron Microscopy-based Nucleic Acid Polymer Sequencing

Project: We aim to provide a comprehensive foundation for development of an ultra-low-cost, ultra-fast nucleic acid polymer sequencing technology based on single-atom resolution transmission electron microscopy (TEM) of heavy atom-labeled nucleic acid polymers.

RC2 HL102815 (NHLBI)

NIH- NHLBI

PI: George Daley (CHB)

Title: Comparative phenotypic, functional, and molecular analysis of ESC and iPSC

9/30/09-8/31/11

Sub: Church: \$76,395

EFFORT

ONRBAA09-001(ONR)

Office of Naval Research

PI: George Church

Title: Multiplexed Pathway and Organism Engineering.

10/1/10- 9/30/11

Church: \$93,835

EFFORT

RC1 HG005482 (NCRR)

NIH - NCRR

PI: Peter Park

Title: Statistical Methods for Estimation of Copy Number from Next – Generation Sequencing

9/22/09-6/30/11

Sub: Church: \$25,756

EFFORT

DOE- DE-AR0000079 (ARPA-E)

ARPA-E

PI: Pamela Silver

Title: Engineering a Bacterial Reverse Fuel Cell

7/1/10-06/30/13

Sub: Church: \$35,494

EFFORT

CBET1033397 (NSF)

PI: Ryan Gill (U. Colorado)

Title: A new approach for directed genome engineering

1/1/11-1/31/13

Sub: Church: \$54,167

EFFORT

DARPA 11-23-CCM-DT-FP-006

PI: Jim Collins (BU)

Title: Synthetic Mammalian Gene Regulatory Circuits for In Vivo Biomedical Applications

6/1/11-5/31/15

Sub: Church: 49,297

EFFORT

ONR

PI: Jim Collins (BU)

Title: Utilizing Synthetic Biology to Create Programmable Micro- Bio- Robots

6/1/11-5/31/16

Sub: Church: \$35,070

EFFORT

Pending:

Pending Support

Form Approved Through 06/30/2012

OMB No. 0925-0001

Department of Health and Human Services
Public Health Services

Review Group

Type

Activity

Grant Number

1

P50

HG005550-02

Grant Progress Report

Total Project Period

From: 09/13/2010

Through: 07/31/2015

Requested Budget Period

From: 08/01/2011

Through: 07/31/2012

1. TITLE OF PROJECT

Causal Transcriptional Consequences of Human Genetic Variation

2a. PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR

(Name and address, street, city, state, zip code)

Kun Zhang
University of California, San Diego
9500 Gilman Drive, Mail Code 0412
La Jolla, California 92093-0412**2b. E-MAIL ADDRESS**

kzhang@eng.ucsd.edu

2c. DEPARTMENT, SERVICE, LABORATORY, OR EQUIVALENT

Bioengineering

2d. MAJOR SUBDIVISION

General Campus

2e. Tel: 858-822-7876

Fax: 858-534-5722

3a. APPLICANT ORGANIZATION

(Name and address, street, city, state, zip code)

The Regents of the Univ. of Calif., San Diego
9500 Gilman Drive, Mail Code 0934
La Jolla, California 92093-0934

3b. Tel: 858-534-0240

Fax: 858-534-0280

3c. DUNS: 804355790

4. ENTITY IDENTIFICATION NUMBER

1956006144A

6. HUMAN SUBJECTS ☐ No ☒ Yes6a. Research
Exempt☒ No ☐ YesIf Exempt ("Yes" in
6a):

Exemption No.

No

If Not Exempt ("No" in
6a):

IRB approval date

08/28/08

5. NAME, TITLE AND ADDRESS OF ADMINISTRATIVE OFFICIALChristine L. Moran, Contract and Grant Officer
Office of Contract and Grant Administration
9500 Gilman Drive, MC0934, La Jolla, CA 92093-0934

Tel: 858-822-2901

Fax: 858-534-0280

E-MAIL: clmoran@ucsd.edu

6b. Federal Wide Assurance No. FWA00004495

6c. NIH-Defined Phase III

Clinical Trial ☒ No ☐ Yes**7. VERTEBRATE ANIMALS** ☒ No ☐ Yes

7a. If "Yes," IACUC approval Date

7b. Animal Welfare Assurance No. A3033-01

10. PROJECT/PERFORMANCE SITE(S)

Organizational Name: The Regents of the Univ. of Calif., SD

DUNS: 804355790

8. COSTS REQUESTED FOR NEXT BUDGET PERIOD

8a. DIRECT \$101,377

8b. TOTAL \$ 148,350

Street 1: 9500 Gilman Drive, Mail Code 0412

Street 2:

9. INVENTIONS AND PATENTS ☒ No ☐ YesIf "Yes," ☐ Previously Reported☐ Not Previously Reported

City: La Jolla

County: San Diego

State: CA

Province:

Country: USA

Zip/Postal Code: 92093-0412

Congressional Districts: CA-053

11. NAME AND TITLE OF OFFICIAL SIGNING FOR APPLICANT ORGANIZATION (Item 13)

Christine L. Moran, Contract and Grant Officer

TEL: 858-822-2901

FAX: 858-534-0280

E-MAIL: clmoran@ucsd.edu

12. Corrections to Page 1 Face Page**13. APPLICANT ORGANIZATION CERTIFICATION AND ACCEPTANCE:** I certify that the statements herein are true, complete and accurate to the best of my knowledge, and accept the obligation to comply with Public Health Services terms and conditions if a grant is awarded as a result of this application. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties.

SIGNATURE OF OFFICIAL NAMED IN

11. (Signature)

DATE

05/09/2011

Cis-regulatory polymorphisms in the human genome.

PI: Kun Zhang, Ph.D.

Over 99% of genetic variations in human population locate outside protein coding sequences. Many phenotypic differences among human individuals are believed to be determined by such non-coding variants. However, identifying causative variants (or mutations) in non-coding regions is very difficult, since they often locate far away from the genes they regulated. Here we propose to use digital RNA allelotyping to identify genes that are regulated by cis-regulatory polymorphisms, and to perform genome-scale haplotyping to establish long-range connectivity between the variants and the genes they regulate.

For digital RNA allelotyping, we will develop and optimize libraries of padlock probes for specific capture of expressed DNA polymorphisms in the human genome. We will couple the padlock capture of human transcriptome with next-generation DNA sequencing to achieve accurate quantification of RNA allelotypes. We will develop a strategy to normalize the capturing efficiency, such that genes with low levels of expression can be quantified efficiently.

For genome-scale haplotyping, we plan to derive several reduced representation genomic libraries from a target genome, each library represents a random subset of chromosomes. Such a polidy conversion procedure will be performed through polymerase cloning on single chromosome molecules. Highly compact metaphase chromosomes will be prepared from primary or immortalized human cell lines. Whole genome amplification will be performed on diluted chromosomes in 8~12 chromosomes per reaction. Each of the resulting amplicon will represent 1/3~1/2 haploid genome. By random shotgun sequencing of 8~10 amplicons, we expect to establish haplotypes for 20~21 chromosomes. To process a large number of Personal Genome samples, we plan to implement the polymerase cloning procedure with microfluidic devices, such as actively controlled nanoreactors or passive microwell arrays. We expect to establish a robust and scalable haplotyping pipeline toward the end of this project.

Program Director/Principal Investigator (Last, First, Middle): Zhang, Kun

Use only if additional space is needed to list additional project/performance sites.

Additional Project/Performance Site Location

Organizational Name: The Regents of the University of California, U.C. San Diego

DUNS: 804355790

Street 1: 9500 Gilman Drive, Mail Code 0412

Street 2:

City: La Jolla

County: San Diego

State: CA

Province:

Country: USA

Zip/Postal Code: 92093-0412

Project/Performance Site Congressional Districts: CA-053

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1:

Street 2:

City:

County:

State:

Province:

Country:

Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1:

Street 2:

City:

County:

State:

Province:

Country:

Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1:

Street 2:

City:

County:

State:

Province:

Country:

Zip/Postal Code:

Project/Performance Site Congressional Districts:

Additional Project/Performance Site Location

Organizational Name:

DUNS:

Street 1:

Street 2:

City:

County:

State:

Province:

Country:

Zip/Postal Code:

Project/Performance Site Congressional Districts:

Program Director/Principal Investigator (Last, First, Middle): ZHANG, Kun

DETAILED BUDGET FOR NEXT BUDGET PERIOD – DIRECT COSTS ONLY	FROM 08/01/2011	THROUGH 07/31/2012	GRANT NUMBER 1P50 HG005550-02
---	---------------------------	------------------------------	---

List PERSONNEL (Applicant organization only)

Use Cal, Acad, or Summer to Enter Months Devoted to Project

Enter Dollar Amounts Requested (omit cents) for Salary Requested and Fringe Benefits

NAME	ROLE ON PROJECT	Cal. Mnths	Acad. Mnths	Summer Mnths	SALARY REQUESTED	FRINGE BENEFITS	TOTALS
Zhang, Kun	PD/PI	EFFORT			5,354	888	6,242
Gole, Jeffrey	Graduate Student				25,629	274	25,903
SUBTOTALS					30,983	1,162	32,145

CONSULTANT COSTS

EQUIPMENT (Itemize)

SUPPLIES (Itemize by category)

Project Specific Costs

49,432

TRAVEL

Domestic Travel for PI to Scientific Meetings

3,000

INPATIENT CARE COSTS

OUTPATIENT CARE COSTS

ALTERATIONS AND RENOVATIONS (Itemize by category)

OTHER EXPENSES (Itemize by category)

Next Generation Network (NGN) Costs \$45 Tuition Remission \$15,255

Publications \$1,500

16,800

SUBTOTAL DIRECT COSTS FOR NEXT BUDGET PERIOD**\$ 101,377**

CONSORTIUM/CONTRACTUAL COSTS

DIRECT COSTS

CONSORTIUM/CONTRACTUAL COSTS

FACILITIES AND ADMINISTRATIVE COSTS

46,973

TOTAL DIRECT COSTS FOR NEXT BUDGET PERIOD (Item 8a, Face Page)**\$ 148,350**

Program Director/Principal Investigator (Last, First, Middle): Zhang, Kun

BUDGET JUSTIFICATIONGRANT NUMBER
1P50 HG005550

Provide a detailed budget justification for those line items and amounts that represent a significant change from that previously recommended. Use continuation pages if necessary.

No significant changes

CURRENT BUDGET PERIODFROM
08/01/2010THROUGH
07/31/2011

Explain any estimated unobligated balance (including prior year carryover) that is greater than 25% of the current year's total budget.

None

Budget Justification (UCSD)

Personnel Total \$ 32,145/year

Kun Zhang, Ph.D. (Principal Investigator, salary requested) **\$6,242/year**

Dr. Kun Zhang will be PI of the UCSD research group. He will oversee the development and optimization methods for RNA allelotyping and full genome haplotyping. He will commit on this project.

Jeff Gole, Graduate student () Stipend **\$25,903/year**

Chromosome preparation, design and fabricate microwell chips, in situ whole genome amplicon on single chromosomes, library construction and shotgun sequencing.

Reagents and supplies Total \$49,432/year

We request an annual budget of \$49,521 to cover Illumina sequencing, Sanger sequencing, library preparation, enzymes, chemicals, tissue culture medium, consumables, SDSC Triton computing cluster fees, user fees for the nanofabrication facility (Nano3) in UCSD.

Travel \$3,000/year

Domestic meetings: Funds are requested so that each researcher may attend a conference each year (total of 2 persons). We estimate that the cost of a conference will be, on average, \$1,500 per person and will cover travel, housing, registration, meals, and incidentals.

Publications \$1,500/year

Funds of \$1,500 is requested to cover the publication cost.

Communication \$45/year

Funds of \$46/year are requested for communication (phone and Ethernet) based on the standard UCSD rate.

Graduate Student Tuition Remission \$15,255/year

Principal Investigator/Program Director (Last, First, Middle): Zhang, Kun

BIOGRAPHICAL SKETCH

NAME Kun Zhang		POSITION TITLE Assistant Professor of Bioengineering	
eRA COMMONS USER NAME eRA Commons User Name			
EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)			
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY
Fudan University, Shanghai, China	B.S.	1996	Biophysics
Fudan University, Shanghai, China	M.S.	1999	Neuroscience
University of Texas-Houston/MD Anderson Cancer Center, TX	Ph. D.	2003	Human & Molecular Genetics
Harvard Medical School, MA	Post-doc	2003-2007	Genetics & Genomics

A. Personal Statement.

Dr. Zhang received his Ph.D. training in the area of human population genetics and cancer genetics in UT-Houston/MD Anderson Cancer Center. During 2003-2007, he was a post-doctoral fellow in the Department of Genetics at Harvard Medical School under the guidance of Dr. George Church. In this period he developed a method for single-cell genome sequencing and another method for long-range haplotyping. In the second phase of his post-doctoral training he started the integration of the padlock capture technology with Agilent's large-scale oligonucleotide synthesis, which led to a number of applications including exome sequencing, RNA allelotyping and target bisulfite sequencing. His current research interest lies in the development of novel genomics approaches and applications to regenerative medicine and human diseases.

B. Positions and Honors.**Positions and Employment**

2000-2002 Rosalie B. Hite Fellow, University of Texas -MD Anderson Cancer Center
 2002-2003 Graduate Research Assistant, Center for Genome Information, University of Cincinnati
 2003-2007 Post-doctoral associate, Department of Genetics, Harvard Medical School
 2007-Present Assistant Professor, Department of Bioengineering, University of California at San Diego

Honors

2000-2002 Rosalie B. Hite Fellowship in Cancer Research, UT- MD Anderson Cancer Center
 2003 Sowell-Huggins Scholarship in Cancer Research, UT- MD Anderson Cancer Center
 2007 Rising Young Investigator, Genome Technology Magazine

C. Selected peer-reviewed publications (in chronological order).**Most relevant to the current application**

1. **Zhang K**, Zhu J, Shendure J, Porreca GJ, Aach JD, Mitra RD, Church GM. (2006) Polony haplotyping of individual human chromosome molecules. **Nature Genetics** 38:382-387.
2. **Zhang K**, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM. (2006) Sequencing genomes from single cells by polymerase cloning. **Nature Biotechnology** 24:680-686.
3. **Zhang K**, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, LeProust EM, Eggan K, Church GM. (2009) Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. **Nature Methods**, 6:613-8
4. Deng J, Shoemaker R, Gore A, Leproust E, Antosiewicz-Gourget, Egli D, Maherli N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y, **Zhang K**. (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. **Nature Biotechnology**, 27:353-60

Principal Investigator/Program Director (Last, First, Middle): Zhang, Kun

5. Gore AJ, Li Z, Fung HL, Young JE, Agarwal S, Antosiewicz-Gourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, Lee JK, Loh YH, Manos PD, Montserrat N, Panopoulos AD, Ruiz S, Wilbert ML, Yu J, Kirkness EF, Belmonte JCI, Rossi DJ, Thomson JA, Eggan K, Daley GQ, Goldstein LSB, **Zhang K. (2011)** Somatic coding mutations in human induced pluripotent stem cells. **Nature** 471:63-7

Additional publications of importance to the field

6. **Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L. (2003)** Randomly distributed crossovers may generate block-like pattern of linkage disequilibrium: An act of genetic drift. **Human Genetics** 113: 51-59.
7. **Zhang K, and Jin L. (2003)** HaploBlockFinder: haplotype block analyses. **Bioinformatics** 19:1300-1301.
8. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, **Zhang K, Mitra RD, Church GM. (2005)** Accurate multiplex polony sequencing of an evolved bacterial genome. **Science** 309:1728-1732.
9. Porreca PJ, **Zhang K***, Li JB, Xie B, Austin D, Vassallo SL, Leproust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J. **(2007)** Multiplex amplification of large sets of human exons. **Nature Methods** 4:931-6.
10. Chen AE, Egli D, Niakan K, Deng J, Akutsu H, Yamaki M, Cowan C, Fitz-Gerald C, **Zhang K, Melton DA, and Eggan K. (2009)** Optimal timing of inner cell mass isolation increases the efficiency of human embryonic stem cell derivation and allows generation of sibling cell lines. **Cell Stem Cell** 4:103-6
11. Li JB, Levanon EY, Yoon J-K, Aach J, Xie B, LeProust E, **Zhang K, Gao Y, Church GM (2009)** Genome-wide Identification of Human RNA Editing Sites by Massively Parallel DNA Capturing and Sequencing. **Science**, 324:1210-3.
12. Lee JH, Park IH, Gao Y, Li JB, Li Z, Daley GQ, **Zhang K***, Church GM*. **(2009)** A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. **PLoS Genetics**, 5(11):e1000718.
13. Shoemaker R, Deng J, Wang W, **Zhang K. (2010)** Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. **Genome Research**, 20:883-9
14. Liu GH, Barkho BZ, Ruiz S, Diep D, Qu J, Yang SL, Panopoulos AD, Suzuki K, Kurian L, Walsh C, Thompson J, Boue S, Fung HL, Sancho-Martinez I, **Zhang K, Iii JY, Belmonte JC. (2011)**. Recapitulation of premature ageing with iPSCs from Hutchinson-Gilford progeria syndrome. **Nature** 472:221-5
15. Howden SE, Gore A, Li Z, Fung HL, Nisler BS, Nie J, Chen G, McIntosh BE, Goulbranson DR, Diol NR, Taapken SM, Vereide DT, Montgomery KD, **Zhang K, Gamm DM, Thomson JA. (2011)** Genetic correction and analysis of induced pluripotent stem cells from a patient with gyrate atrophy. **PNAS** [PMID: 21464322]

D. Research Support

Current Research Support

NIH - NIDA R01 DA025779 2008-2012

Contact PI: Kun Zhang

Title: Genome-scale analysis of DNA methylation in CpG islands through bisulfite sequencing.

NIH - NHGRI R01 HG 004876 2008-2011

Contact PI: Kun Zhang

Title: An integrated lab-on-chip system for genome sequencing of single microbial cells.

NIH - NIGMS R01 GM 097253 2011-2015

PI: Kun Zhang

Title: Genome-methylome interaction

NIH/NHGR P50 HG005550 (Church) 2010-2015

Principal Investigator/Program Director (Last, First, Middle): Zhang, Kun

Role on project: Co-PI

Title: Causal Transcriptional Consequences of Human Genetic Variation

NSF OCE **10463683 (Martiny)**

2011-2016

Co-PI: Kun Zhang

Title: Dimensions: Collaborative research: Biological controls of the ocean C:N:P ratios.

Program Director/Principal Investigator (Last, First, Middle): Zhang, Kun

PROGRESS REPORT SUMMARY	GRANT NUMBER P50 HG005550	
	PERIOD COVERED BY THIS REPORT	
PROGRAM DIRECTOR / PRINCIPAL INVESTIGATOR Kun Zhang	FROM 9/13/10	THROUGH 5/31/11
APPLICANT ORGANIZATION UCSD		
TITLE OF PROJECT (Repeat title shown in Item 1 on first page) Causal Transcriptional Consequences of Human Genetic variation		
A. Human Subjects (Complete Item 6 on the Face Page)		
Involvement of Human Subjects	<input checked="" type="checkbox"/> No Change Since Previous Submission	<input type="checkbox"/> Change
B. Vertebrate Animals (Complete Item 7 on the Face Page)		
Use of Vertebrate Animals	<input checked="" type="checkbox"/> No Change Since Previous Submission	<input type="checkbox"/> Change
C. Select Agent Research	<input checked="" type="checkbox"/> No Change Since Previous Submission	<input type="checkbox"/> Change
D. Multiple PD/PI Leadership Plan	<input checked="" type="checkbox"/> No Change Since Previous Submission	<input type="checkbox"/> Change
E. Human Embryonic Stem Cell Line(s) Used	<input checked="" type="checkbox"/> No Change Since Previous Submission	<input type="checkbox"/> Change

SEE PHS 2590 INSTRUCTIONS.

WOMEN AND MINORITY INCLUSION: See PHS 398 Instructions. Use Inclusion Enrollment Report Format Page and, if necessary, Targeted/Planned Enrollment Format Page.

Program Director/Principal Investigator (Last, First, Middle): Church, George M, Zhang Lag CCV Progress Report 2011

Zhang Lag CCV Progress Report 2011

Improvement of ASE assay. We previously designed two sets of padlock probes for detecting allele-specific gene expression (ASE) in the human genome, and demonstrated that ASE is both common, tissue specific, and a strong indicator for the presence of cis-regulatory variants (Lee et al. 2009; Zhang et al. 2009). There were two remaining technical limitations: the capture efficiencies were still far from even, and the presence of allele-specific capture bias. We recently designed and synthesized a third generation of probe set with significant improvements on both aspects. We used a neural network model that was trained with experimental data to design probes with a lower variation in capture efficiency. In addition, the gap size of the padlock capturing arms was increased to 18nt, such that the variant sites locate outside the footprint of DNA ligase and away from the extension terminal of DNA polymerase. Finally, we also increased the size of probe set from 27,000 to 36,000 coding SNPs. With 6-8 million Illumina 36bp sequencing reads, we were able to confidently call ~30,000 SNPs on genomic DNA and 22,000 expressed SNPs on cDNA, among which roughly 5,000 are heterozygous informative SNPs. We can multiplex at least 30 samples in one Illumina GA IIx sequencing flowcell, and 100 samples in HiSeq sequencing flowcell, making it a very efficient and cost effective assay for ASE. We are in the process of performing a more complete survey of ASE in the three cell types (EB-transformed lymphocytes, primary fibroblasts, induced pluripotent stem cells) derived from PGP1, by combining padlock capture with this third generation probe set with high-coverage RNAseq.

Allele-specific methylation (ASM) analysis. We have recently discovered that the presence of cis-regulatory genetic variants not only leads to allele-specific gene expression (ASE), it also results in differential epigenetic modifications on the two parental chromosomes, such as allele-specific methylation (ASM). In a recent study focusing on all CpG islands on two human chromosomes, we showed that ASM is present in the human genome at a similar level as ASE, and one mechanism through which genetic polymorphisms interact with the epigenome is direct creation or removal of CpG dinucleotides by genetic variants (Shoemaker et al. 2010). More recently, we have further expanded our survey of ASM to all 23 human chromosomes. Using a library of 330,000 bisulfite padlock probes, we quantified over 500,000 CpG sites in known differentially methylated regions and DNase I hypersensitive, and discovered thousands of ASM regions in the three PGP1 cell lines as well as additional EBV-transformed lymphocytes derived from other PGP donors and HapMap individuals (Table 1). We will integrate ASM and ASE events identified in the same set of samples to further understand how cis-regulatory variants regulate both the epigenome and transcriptome.

Whole genome haplotyping. We proposed to connect cis-regulatory genetic variants with their regulating genes by building complete diploid genome maps in which all genetic variants on 23 chromosomes are fully phased. Our strategy for generating a fully phased diploid genome involved preparation of metaphase chromosome molecules, followed by whole genome amplification of random subsets of human chromosomes, and shotgun sequencing. We have gone through one complete cycle of this experiment, and identified two sequencing libraries that were derived from a small number of human chromosomes. As shown in Figure 1, these results are promising in that single intact human chromosomes can be prepared from metaphase trapped cells and be amplified relatively even across the chromosomes; in addition, variants called on the three intact chromosomes are mostly homozygous, indicating that only one particular parental allele was present. There are two remaining issues: (i) the presence of sequencing reads from other random chromosomal locations; (ii) relative sparse sequencing coverage on the intact chromosomes, which is similar to what was recently reported with microfluidic-base MDA (Fan et al. 2011). The first issue is mostly likely due to the presence of small chromosomal fragments in the amplification templates. A simple solution is to use FACS sorting to separate intact chromosomes from small DNA fragments that were generated as the side-products of chromosome preparation. We are in the process of optimizing protocols for

Program Director/Principal Investigator (Last, First, Middle): Church, George M, Zhang Lag CCV Progress Report 2011

chromosome sorting. To improve the sequencing coverage, we are tested three different strategies, including (i) tuning down the yield of whole genome amplification (to reduce amplification bias) coupled with developing library-construction protocols that require low-input DNA; (ii) increasing the number of amplicons for each sample in order for upsampling of chromosomes; (iii) significantly increase the amount of sequencing per sample. We are optimistic that the diploid genome of PGP1 will be completely phased within one year since the start of this project (by October 2011).

Table 1

Sample	# raw reads	#CpGs called	#LD blocks	#SNPs called	# het SNPs called	#ASM I ⁺	#ASM II ⁺	#ASM III ⁺
PGP1L	22,134,386	461,908	12,437	17,240	8,787	1,652	167	1,042
PGP1F*	11,894,560	204,582	4,253	8,651	4,717	768	94	507
PGP1iPS*	14,982,406	201,255	4,589	8,641	4,455	407	123	911
PGP2L	28,495,192	543,140	11,153	19,767	10,116	1,131	191	1,545
PGP3L	29,570,157	521,279	15,896	19,235	9,884	1,757	198	1,205
PGP5L	11,576,285	308,050	7,588	12,355	6,350	643	92	882
PGP7L	11,970,974	316,178	8,347	11,917	6,138	938	130	704
PGP8L	26,422,937	484,298	12,069	17,038	8,429	1,359	210	1,208
PGP9L	29,867,962	523,066	11,602	19,164	10,057	979	181	1,660
PGP10L	33,157,508	554,246	13,504	22,369	12,045	2,033	243	1,714
GM12878	26,669,930	507,086	11,765	18,388	9,873	1,948	198	953

Notes: * PGP1F and PGP1iPS were captured with an earlier version of 220k probe set, therefore fewer CpG sites were analyzed. +Allele Specific Methylation (ASM) was divided into three classes: (I) ASM regions that are completely independent of the presence of CpG-SNPs; (II) ASM regions that contain CpG SNPs in the dbsnp131 database, but were not called heterozygous; (III) ASM regions that are due to the presence of heterozygous CpG-SNPs

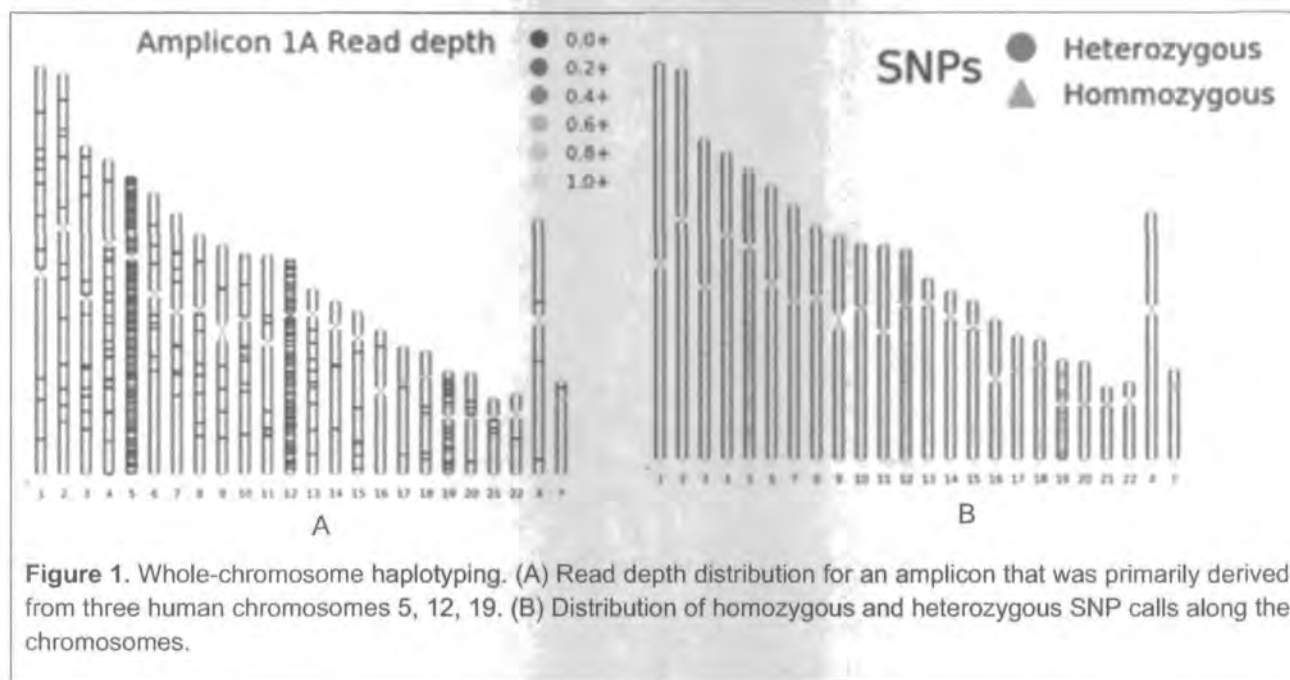


Figure 1. Whole-chromosome haplotyping. (A) Read depth distribution for an amplicon that was primarily derived from three human chromosomes 5, 12, 19. (B) Distribution of homozygous and heterozygous SNP calls along the chromosomes.

References:

Program Director/Principal Investigator (Last, First, Middle): Church, George M, **Zhang Lag CCV Progress Report 2011**

- Fan HC, Wang J, Potanina A, Quake SR. 2011. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* **29**(1): 51-57.
- Lee JH, Park IH, Gao Y, Li JB, Li Z, Daley GQ, Zhang K, Church GM. 2009. A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet* **5**(11): e1000718.
- Shoemaker R, Deng J, Wang W, Zhang K. 2010. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res*.
- Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, Leproust EM et al. 2009. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* **6**(8): 613-618.